Actes des 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT).

Thierry Poibeau, Yannick Parmentier, Emmanuel Schang

LIFT 2020

Actes des 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)

Thierry Poibeau, Yannick Parmentier, Emmanuel Schang (Éds.)

10-11 décembre 2020
Montrouge, France (Virtuel)

Comités

Comité d'organisation

Thierry Poibeau, LATTICE/CNRS

Yannick Parmentier, LORIA/Université de Lorraine

Emmanuel Schang, LLL/Université d'Orléans

Comité de programme

Angélique Amelot LPP/CNRS

Laurent Besacier LIG/Université Grenoble Alpes

Berthold Crysmann LLF/Université de Paris

Karën Fort LORIA/Université Paris Sorbonne

Claire Gardent LORIA/CNRS

Anaïs Lefeuvre-Halftermeyer LIFO/Université d'Orléans

Alexis Michaud LaCITO/INALCO

Tatiana Nikitina LLACAN/CNRS

Yannick Parmentier LORIA/Université de Lorraine

Thierry Poibeau LATTICE/ENS Paris

Emmanuel Schang LLL/Université de Lorraine

Valentin Vydrine LLACAN/INALCO

Guillaume Wisniewski LLF/Université de Paris

Présentations invitées

Katja Aplonova (LLACAN)

Titre : L'annotation syntaxique pour le Bambara

Résumé : La présentation est consacrée aux premières étapes du développement du corpus syntaxique (treebank) du Bambara. Le bambara est une langue mandingue parlée au Mali qui possède d'un corpus avec l'annotation morphologique. Le schéma d'annotation syntaxique est basé sur le modèle Universal Dependencies, un projet fournissant une annotation syntaxique inter-linguistique pour différentes langues. Dans la présentation, les problèmes de conversion de données sont abordés avec une attention particulière au choix des étiquettes des parties de discours et des relations syntaxiques.

Lionel Clément (LaBRI)

Titre : XLFG

Résumé : XLFG est un logiciel d'analyse syntaxique pour grammaires lexicales fonctionnelles (LFG - Bresnan et al. 1982). LFG offre une analyse de la langue qui ne se base ni sur une relation biunivoque entre une structure profonde et une structure de surface, ni sur une construction combinatoire d'éléments atomiques. Les analyses LFG sont des structures syntagmatiques, des structures fonctionnelles, et des structures argumentales. Elles offrent donc trois niveaux d'analyse syntaxique relativement indépendants. L'intérêt principal d'XLFG est d'analyser en temps optimal des phrases complexes avec une grammaire LFG, et de fournir si nécessaire les raisons de l'agrammaticalité de certaines phrases. Par exemple en détectant des ruptures de contraintes d'accord, de valence, de colocation, ou de projection lexicale. Une interface riche (xlfg.org) a été proposée en ligne qui permet aux chercheurs, étudiants et enseignants de créer des ateliers d'écriture de grammaires dans différentes langues. Dans cet exposé, nous expliquerons rapidement comment nous avons résolu l'analyse de phrases dont l'ambiguïté rend inopérante toute tentative classique. Cela passe nécessairement par la définition d'une nouvelle sémantique du système formel de LFG. Nous n'aurons pas le temps de détailler ceci, mais en revanche, nous montrerons l'application par cette sémantique de quelques hypothèses théoriques, dont l'analyse de coordonnées elliptiques en français ou l'analyse des particules verbales en anglais.

Annie Foret (IRISA)

Titre : On Categorial Grammatical Inference and Logical Information Systems.

Résumé : We shall consider several classes of categorial grammars and discuss their learnability. We consider learning as a symbolic issue in an unsupervised setting, from raw or from structured data and treebanks for some variants of Lambek grammars and of categorial dependency grammars. In that perspective, we discuss for these frameworks different type constructors and structures, some limitations (negative results) but also some algorithms (positive results) under some hypothesis. On the experimental side, we also consider the Logical Information Systems approach, that allows for navigation, querying, updating, and analysis of heterogeneous data collections where data are given (logical) descriptors. Categorial grammars can be seen as a particular case of Logical Information System.

Kim Gerdes (LIMSI)

Titre : Un treebank dépendanciel du Naija

Résumé : Le développement d'un treebank du Naija parlé (le pidgincréole du Nigéria) dans le cadre du projet ANR NaijaSynCor http ://naijasyncor.huma-num.fr/ joue un rôle dans les enjeux de linguistique politiques du pays, mais il est aussi intéressant d'un point de vue tal et linguistique : avec ces 475k mots, le treebank est le plus grand treebank oral existant et son développement constituait des défis tal intéressant. La qualité et la cohérence du treebank étaient assurées par un entraînement régulier d'un parser neuronal (bootstrapping), par des corrections globales à l'aide des grammaires de réécriture de graphe (Grew) et par des extractions, corrections et réintroductions récurrentes de lexiques. Actuellement, nous travaillons au développement d'un dictionnaire collaboratif (wiktionnaire) dont la structure et des exemples sont extraits directement du treebank.

Sylvain Loiseau (LACITO)

Titre : Les gloses interlinéaires : de la description au corpus en typologie linguistique

Résumé : Les gloses interlinéaires sont un format pour représenter l'analyse morphosyntaxique d'un texte ou d'un énoncé. Ce format joue un rôle central dans la documentation et la description des langues du monde ainsi qu'en typologie linguistique. Il est destiné à l'origine à représenter une analyse pour justifier une argumentation plus qu'à être une structure de données permettant des traitements automatiques. Cependant de vastes corpus de textes annotés dans ce format sont aujourd'hui disponibles. L'écosystème pour le traitement, la publication ou l'analyse automatique de ces données est encore peu développé. Cette présentation discutera des possibilités offertes par ces données, des outils disponible pour les analyser de façon plus systématique, et enfin des briques logicielles manquantes pour les faire accéder à des méthodologies de type linguistique de corpus (notamment en termes de modélisation quantitative).

Aleksandra Miletic (CLLE)

Titre : Construire un treebank pour une langue peu dotée : ce que nous apprenent les cas du serbe et de l'occitan

Résumé : Je propose de partager mon expérience sur deux campagnes de création de treebank que j'ai menées jusqu'ici, la première pour le serbe et la deuxième pour l'occitan. Les deux langues étant peu dotées en ressources du TAL librement disponibles au moment de démarrage, ces projets ont été en grande partie basés sur l'annotation manuelle. Dans un tel contexte, comment rendre le travail des annotateurs humains aussi facile et rapide que possible, tout en préservant la qualité des annotations produites ? Est-il utile de chercher à exploiter les ressources existantes pour des langues proches ? Vaut-il mieux définir un schéma d'annotation propre à la langue en question ou en adopter un déjà existant ? Je présente les solutions mises en place dans chacune des campagnes, leurs résultats, mais aussi les questionnements qu'elles ont suscités.

Table des matières

ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs – Nouvelles fonctionnalités

Emmanuelle Esperança-Rodier[1] Francis Brunet-Manquat[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France

emmanuelle.esperanca-rodier@univ-grenoble-alpes.fr, francis.brunet-manquat@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article présente les avancées d'ACCOLÉ (Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs), qui en plus de proposer une gestion simplifiée des corpus et des typologies d'erreurs, l'annotation d'erreurs pour des corpus de traduction bilingues alignés, la collaboration et/ou supervision lors de l'annotation, la recherche de modèle d'erreurs dans les annotations, permet désormais d'annoter les Expressions Polylexicales (EPL) dans des textes monolingues en français, et d'accéder à l'annotation d'erreurs pour des corpus de traduction multi-cibles. Dans cet article, après un bref rappel des fonctionnalités d'ACCOLÉ, nous explicitons les fonctionnalités de chaque nouveauté.

ABSTRACT

ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora – New Features

This article presents the recent advances in ACCOLÉ, which on top of offering simplified management of corpora and typologies of errors, annotation of errors in bilingual aligned corpora, collaboration and/or supervision during annotation, looking for error types in annotations, now permits to annotate Multi-Word Expressions (MWE) in French monolingual corpora, and to access error annotation for multi-target corpora. In this article, after reminding the regular features of ACCOLÉ, we will explain the features of each novelty.

MOTS-CLÉS : Annotations d'erreurs de Traductions Automatique, Annotation collaborative, Evaluation de la qualité de la TA, EPL

KEYWORDS: Annotations of translation errors, Collaborative annotation, Machine Translation Quality Assessment, MWE

1 ACCOLÉ, une plateforme pour l'annotation d'erreurs

ACCOLÉ permet l'annotation manuelle des erreurs de traduction selon des critères linguistiques. L'idée sous-jacente est de pouvoir fournir à un utilisateur une aide dans le choix d'un système de TA à utiliser selon le contexte (compétences linguistiques et informatiques de l'utilisateur, connaissance du domaine du document source à traduire et la tâche pour laquelle il a besoin de traduire le document source.) Pour ce faire, ACCOLÉ doit permettre de détecter quels sont les phénomènes linguistiques qui ne sont pas traités correctement par le système de TA étudié. Les principales fonctionnalités de la plateforme ACCOLÉ sont la gestion simplifiée des corpus, des typologies d'erreurs, des annotateurs, etc. ; l'annotation d'erreurs ; la collaboration et/ou supervision lors de l'annotation ; la recherche de modèles d'erreurs (type d'erreurs dans un premier temps, patrons morphosyntaxiques ultérieurement) dans les annotations. Nous avons privilégié un accès simple à l'outil ainsi qu'au corpus. La plateforme ACCOLÉ est donc disponible en ligne (http://lig-accole.imag.fr, documentation : http://lig-membres.imag.fr/fbrunet/accole-plateforme-pour-ledition-collaborative-derreurs/) depuis un navigateur et ne nécessite aucune installation spécifique.

1.1 Gestion des projets d'annotations

Un projet d'annotation renvoie à une tâche d'annotation, en créant un couple associant un corpus et une typologie d'annotation. Ainsi, un même corpus pourra être associé à plusieurs typologies sous forme de plusieurs projets d'annotation. Ainsi, le corpus ne sera chargé qu'une fois sur la plateforme. Les annotateurs ainsi que les superviseurs sont associés aux projets qu'ils doivent annoter par le responsable du projet. Les typologies d'erreur sont également gérées par le responsable du projet. Un type d'erreur est composé d'un nom, d'une catégorie (facultative), d'une sous-catégorie (facultative) et d'un code (raccourci clavier pouvant être utilisé lors de l'annotation). Des typologies déjà existantes, telles que Vilar et al. (2006) et DQF-MQM (Lommel et al., 2018) sont décrites de cette manière dans ACCOLÉ. Des corpus (français-anglais) sont déjà disponibles dans ACCOLÉ, comme des extraits du BTEC (Basic Travel Expression Corpus), des nouvelles journalistiques, des articles sur la réglementation européenne issus de l'Union Européenne, des brevets. De plus, la plate-forme permet de téléverser de nouveaux corpus et de saisir d'autres typologies d'erreurs.

1.2 Annotation d'erreurs

La plateforme ACCOLÉ propose de visualiser et d'annoter les erreurs d'un couple de phrases source/cible (mono-cible) (Figure 1).

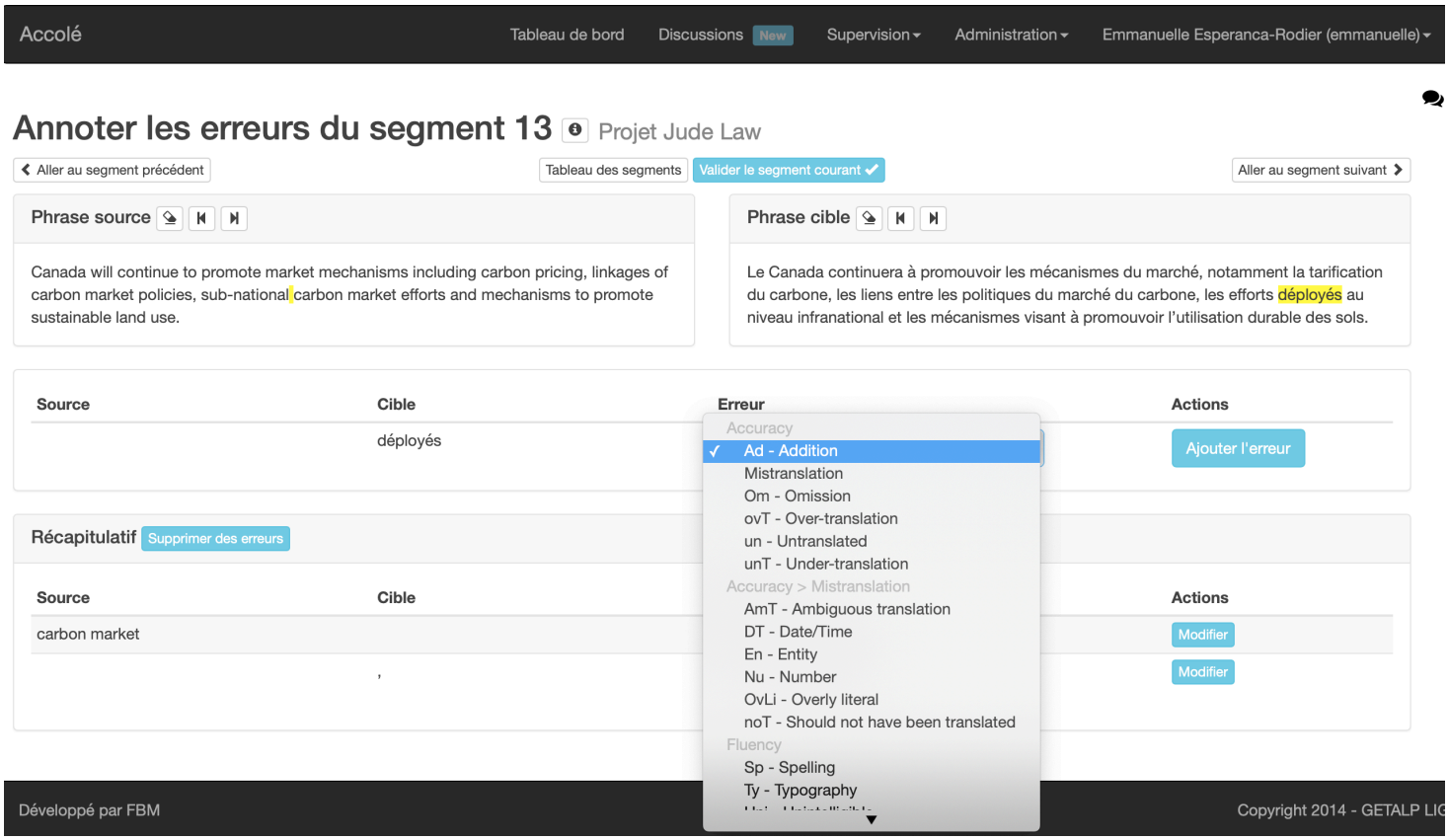

FIGURE 1 : Annotation d'une erreur sur la plateforme ACCOLÉ avec la typologie DQF-MQM (Lommel et al., 2018) dans un corpus mono-cible.

La figure 1 présente l'interface pour un corpus mono-cible proposée à l'annotateur. L'annotation se fait en deux étapes. La première étape consiste à sélectionner, à l'aide de la souris, des mots dans la phrase source, et de leur équivalent dans la phrase cible, présentant une erreur de traduction. Il est possible de sélectionner des mots disjoints dans la source et dans la cible. Dans le cas de mots non traduits (omission), il faut sélectionner l'espace dans la cible, à l'endroit où le ou les mots sources aurait dû être traduits. Dans le cas d'addition, il faut sélectionner l'espace entre les mots sources,

correspondant à la position du ou des mots qui ont été ajoutés dans la cible entre les traductions de ces mots sources. La seconde étape consiste à choisir le type d'erreurs soit à l'aide de la souris, soit à l'aide des raccourcis clavier, à associer au couple des mots sources/cibles préalablement sélectionnés. Dans la figure 1, nous avons un exemple de mots ajoutés dans la traduction. « sub-national carbon market efforts » a été traduit par *« les efforts déployés au niveau infranational ». « déployés » a été ajouté dans la traduction alors qu'il n'est pas présent dans la source. L'espace entre « sub-national » et « carbon » a donc été sélectionné comme l'endroit équivalent à l'ajout dans la traduction de « déployés ». Cette façon d'annoter facilite la recherche de patrons pour un type d'erreurs, car il permet d'indiquer le contexte d'insertion, ou de suppression du mot respectivement dans la source et la cible.

Pour répondre aux problèmes d'accord inter-annotateurs (Popović, 2018), ACCOLÉ propose deux mécanismes pour aider l'annotateur dans sa tâche. Le premier est un mécanisme de supervision permettant à un responsable de contrôler l'avancée de la tâche. Ce mécanisme encourage surtout la communication entre superviseur et annotateur par la possibilité de créer des fils de discussion pour un couple de phrase source/cible précis. La supervision autorise la demande de précisions sur un type d'erreurs, de pointer une erreur d'annotation, etc.. Le second mécanisme est le mécanisme collaboratif qui permet aux annotateurs de communiquer autour d'un couple phrase source/cible précis. Ce mécanisme est une option à activer dans le projet.

1.3 Représentation des erreurs basée sur les SSTC

La plateforme utilise une représentation des données basée sur les SSTC (Structured String-Tree Correspondences, Boitet et Zaharin 1988). Une erreur est constituée d'une étiquette et d'un ensemble de SNODE (intervalle représentant la sous-chaîne dans la phrase source ou cible correspondante). Par exemple dans la figure 2, l'erreur portant sur "toute l'" et "any" est décrite par son étiquette Mauvais choix lexical (cat. Mot incorrect, sous-cat. Sens), par son positionnement dans la phrase source (SNODE [49-56] - sous chaîne entre le 49ème caractères et le 56ème) et la phrase cible (SNODE [46-48]). L'avantage d'utiliser ainsi les SNODE est de se passer d'une structure syntaxique pour décrire l'erreur.

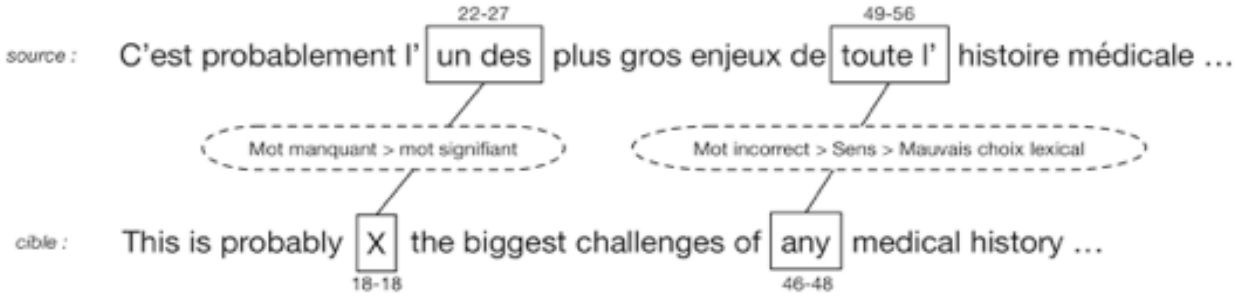

FIGURE 2 : exemple d'annotations

L'autre avantage est de pouvoir ajouter a posteriori des analyses morphosyntaxiques sur les phrases sources et cibles. Une ou plusieurs analyses (Stanford Tagger, Xerox Incremental Parser, etc.) pourront ainsi être rattachées à l'aide des SNODE aux phrases. A la fin de la tâche d'annotation, ses analyses serviront à rechercher des modèles d'erreurs (patrons morphosyntaxiques par exemple). L'idée est donc d'utiliser les erreurs comme une représentation pivot dans le mécanisme de recherche.

2 ACCOLÉ, nouvelles fonctionnalités

2.1 Annotation d'erreurs multi-cibles

Dans le cadre d'une expérimentation sur l'analyse de la qualité de systèmes de Traduction Automatique Neuronale en traduction simultanée ou après complétion de la phrase entière - online & offline NMT - (Elbayad et al., 2020), ACCOLÉ s'est étoffée de la possibilité d'annoter plusieurs hypothèses de traduction correspondant à une seule phrase source et d'intégrer également une phrase de référence (Figure 3). Cette étude porte sur l'évaluation de la qualité des hypothèses de traduction, d'une même phrase source, issues de quatre systèmes différents de TA neuronale. Il s'agit d'effectuer une annotation des erreurs pour les quatre hypothèses de traduction en langue cible par rapport à la phrase source et à une référence de traduction. La typologie utilisée est une version adaptée de MQM-DQF (Lommel et al., 2018). Le corpus anglais-allemand annoté dans ce projet correspond à 200 segments issus des données IWSLT14.

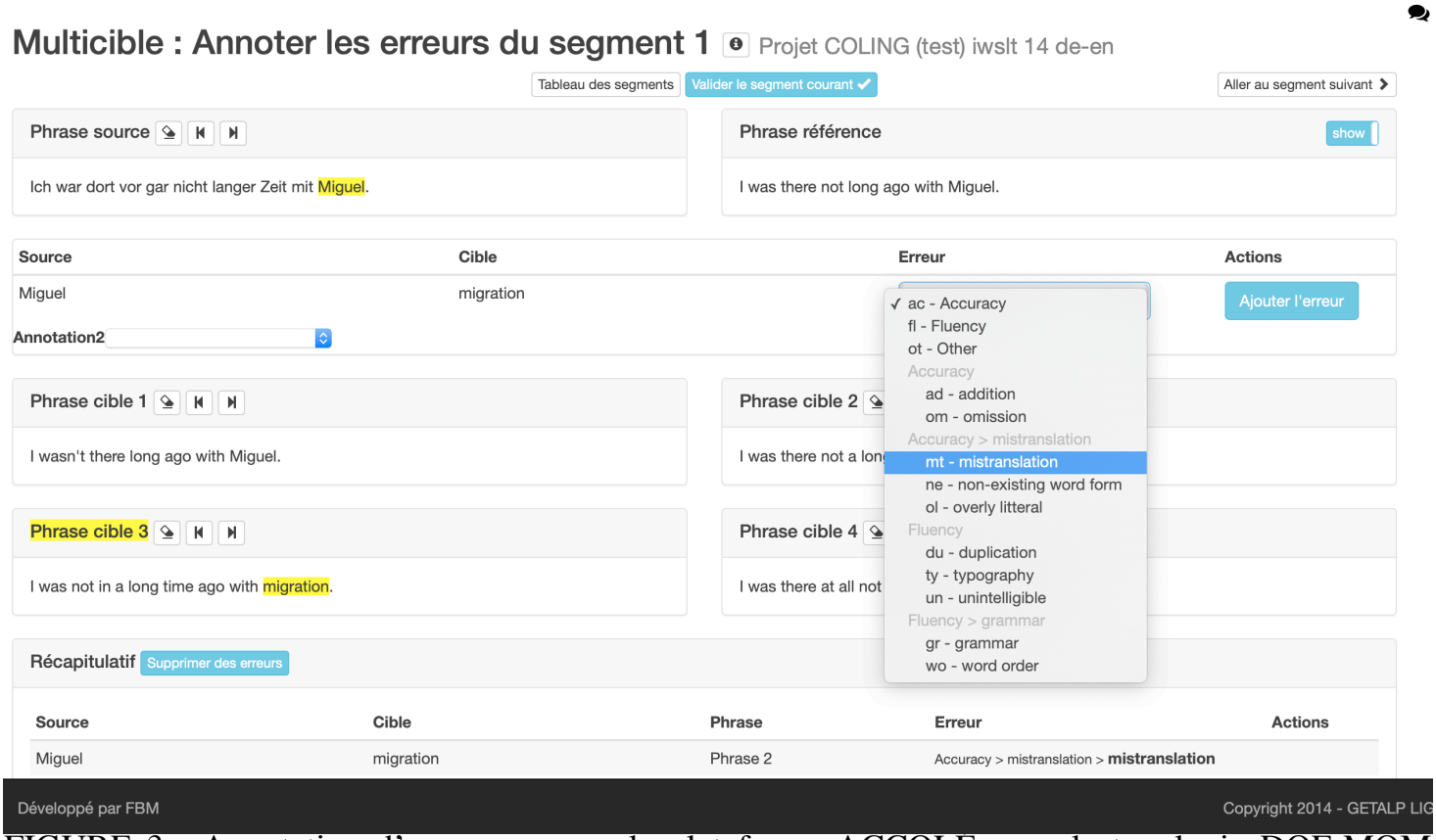

FIGURE 3 : Annotation d'une erreur sur la plateforme ACCOLÉ avec la typologie DQF-MQM (Lommel et al., 2018) dans un corpus multi-cible avec référence.

L'annotation d'erreurs sur un corpus multi-cibles se déroule de la même façon que pour un corpus mono-cible. L'annotateur sélectionne à l'aide de la souris le couple d'occurrence source/cible 1 source/cible 2, source /cible 3… présentant une erreur de traduction. La seconde étape consiste à choisir le type d'erreurs soit à l'aide de la souris, soit à l'aide des raccourcis clavier, à associer au couple des mots sources/cibles préalablement sélectionnés. En plus de la source, l'annotateur a accès à une traduction de référence. Dans l'exemple de la figure 3, « Miguel » a été traduit par « Miguel » dans la cible 1et la cible 4, et par *« migration » dans la cible 2 et la cible 3. Une erreur est donc annotée entre « Miguel » de la source, et *« migration » dans la cible 2 avec pour type « Accuracy>Mistranslation ». Comme le troisième moteur de traduction a engendré la même erreur de traduction, cette dernière est également annotée pour la cible 3 avec le même type d'erreur

2.2 Annotation d'Expressions Polylexicales

Dans le cadre d'un projet financé par NeuroCoG/Pôle Grenoble Cognition, nous avons adapté ACCOLÉ pour l'annotation d'Expressions Polylexicales (EPL). La typologie de types d'EPL intégrée à notre plateforme est telle que définie dans le travail de Tutin & Esperança-Rodier (2017). Elle est composée de 9 types : Collocations (C), Mots Fonctionnels (F), Formules de Routine (FR), Entités nommées (NE), Phrasèmes complets (PH), Pragmatèmes (PRAG), Proverbes (PROV),

Collocations fortes (SC) et enfin Termes Complexes (T). Chaque EPL est également annoté en partie du discours.

Nous avons envisagé que le corpus annoté soit monolingue ou bien bilingue. Toutefois, nous préférons la possibilité d'annoter la source en EPL de manière monolingue, de même que la cible. Si le corpus possède une traduction alignée du texte, alors il est possible d'annoter, à la fois dans la source et dans la cible, l'erreur repérée entre une EPL et sa traduction, afin de faire correspondre et comparer les annotations faites en première étape monolingue. Un projet financé par NeuroCoG/Pôle Grenoble Cognition, débutant en janvier 2021, nous permettra de vérifier de facto ce qui est le plus adapté pour l'annotateur.

Afin de faciliter la tâche d'annotation, un dictionnaire monolingue français d'EPL a été ajouté à ACCOLÉ, ainsi qu'un pré-traitement basé sur l'analyse syntaxique (Coavoux et Crabbé, 2017). Ainsi, ACCOLÉ permet d'annoter des EPL soit manuellement, en sélectionnant des mots à l'aide de la souris et en leur assignant un type, soit sur proposition de l'interface utilisant de manière automatique le dictionnaire et le pré-traitement, proposition qui sera à valider par l'annotateur. Dans un projet, on peut choisir d'obtenir une proposition automatique des annotations (Figure 4) ou non.

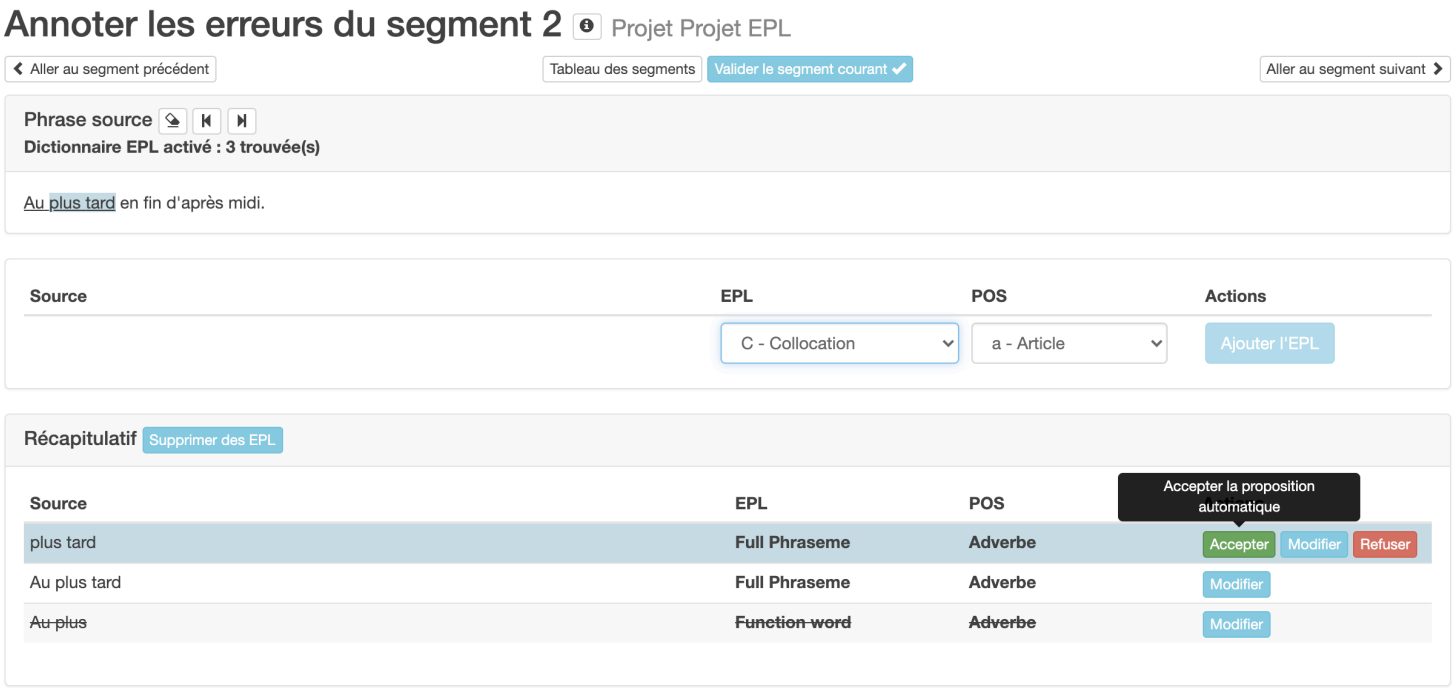

Figure 4 : Annotation d'un segment avec proposition automatique d'annotations à partir du dictionnaire

Dans la pratique, le dictionnaire est :

- soit utilisé par le mécanisme automatique pour déterminer toutes les EPLs d'un segment source et les proposer à l'annotateur dans l'éditeur (Figure 4). L'annotateur pourra les accepter, les refuser ou les modifier, c'est-à-dire les accepter mais en modifiant le type et/ou la partie du discours de l'EPL.

- soit le projet ne propose pas le chargement automatique des EPLs, mais l'interface d'édition précisera le nombre d'EPLs possibles (voir Figure 5), que l'annotateur décidera d'ajouter ou non.

Le mécanisme automatique repère les EPLs présentes dans le segment source à partir du dictionnaire. Le dictionnaire contient le mot du début d'une EPL. L'algorithme automatique cherche ensuite si les mots restants de l'EPL correspondent à la suite des mots du segment source. L'algorithme est pour le moment très simple car le dictionnaire est très simple. Après la fin des premières annotations, le dictionnaire sera amélioré (factorisation, lemmatisation, lien avec le format NIF (Hellman et al., 2013)). Il prendra en compte ces modifications en introduisant l'utilisation d'un lemmatiseur et, à plus long terme, proposera des annotations avec plus d'informations provenant du web sémantique.

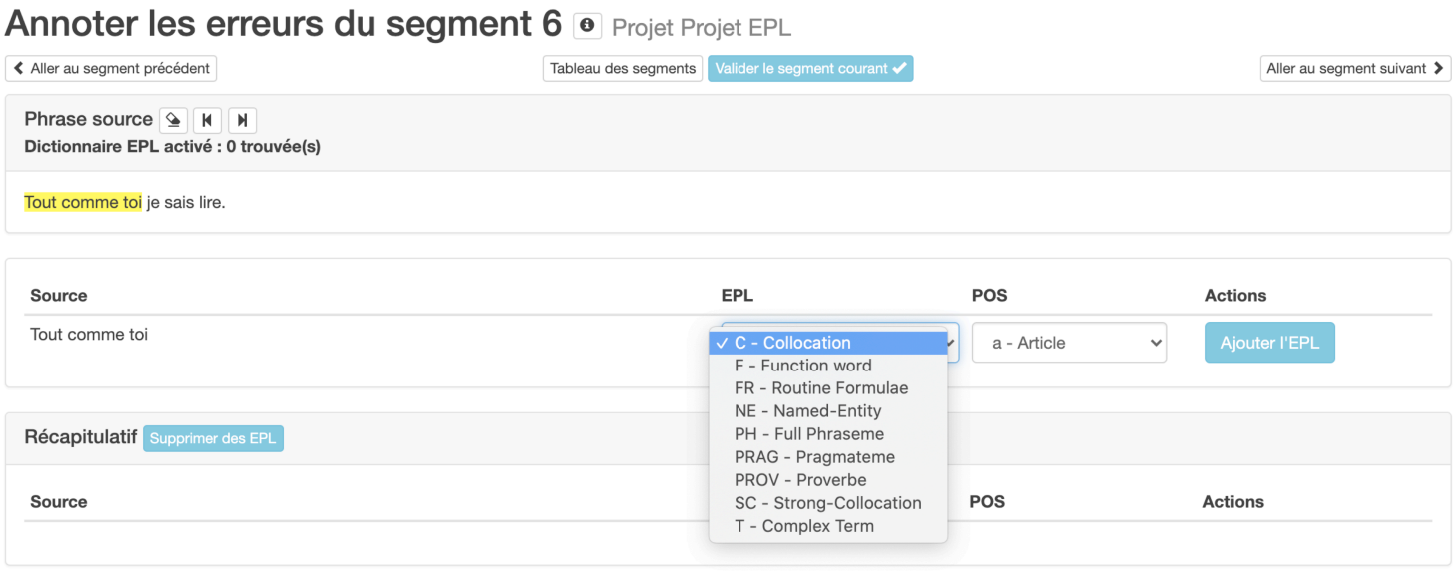

Figure 5 : Annotation d'une EPL de manière manuelle

L'analyse pré-syntaxique, quant à elle, propose déjà l'annotation en EPL du segment source. L'intégration à ACCOLÉ, dans un premier temps, consiste à faire correspondre l'analyse au formalisme d'ACCOLÉ, c'est-à-dire aux SNODEs. Puis dans un second temps, il s'agit d'ajouter les propositions d'EPL pour le segment source, et ce en faisant correspondre les types d'EPL avec les types utilisés pour l'annotation dans le projet et utilisés dans le dictionnaire.

La proposition d'annotation pour le segment source dans l'éditeur graphique sera la même que pour les propositions provenant du dictionnaire. Ces deux mécanismes ne diffèrent pas techniquement mais diffèrent dans le « workflow » de l'édition pour l'annotateur.

Nous avons obtenu un financement de NeuroCoG/Pôle Grenoble Cognition, pour réaliser avec ACCOLÉ l'annotation d'EPL dans un corpus monolingue français de 40,000 mots, que nous ferons traduire en polonais par différents moteurs de traduction automatique, notamment DeepL et Google. Nous étudierons ains le comportement de ces moteurs face à la traduction d'EPL du français vers le polonais. Cette étude nous permettra d'agrandir la quantité de corpus à disposition dans la plateforme ACCOLÉ, ainsi que d'ajouter le coupe français-polonais aux couples de langues jusque là étudiés. Nous serons également en mesure de créer un dictionnaire d'EPL en polonais.

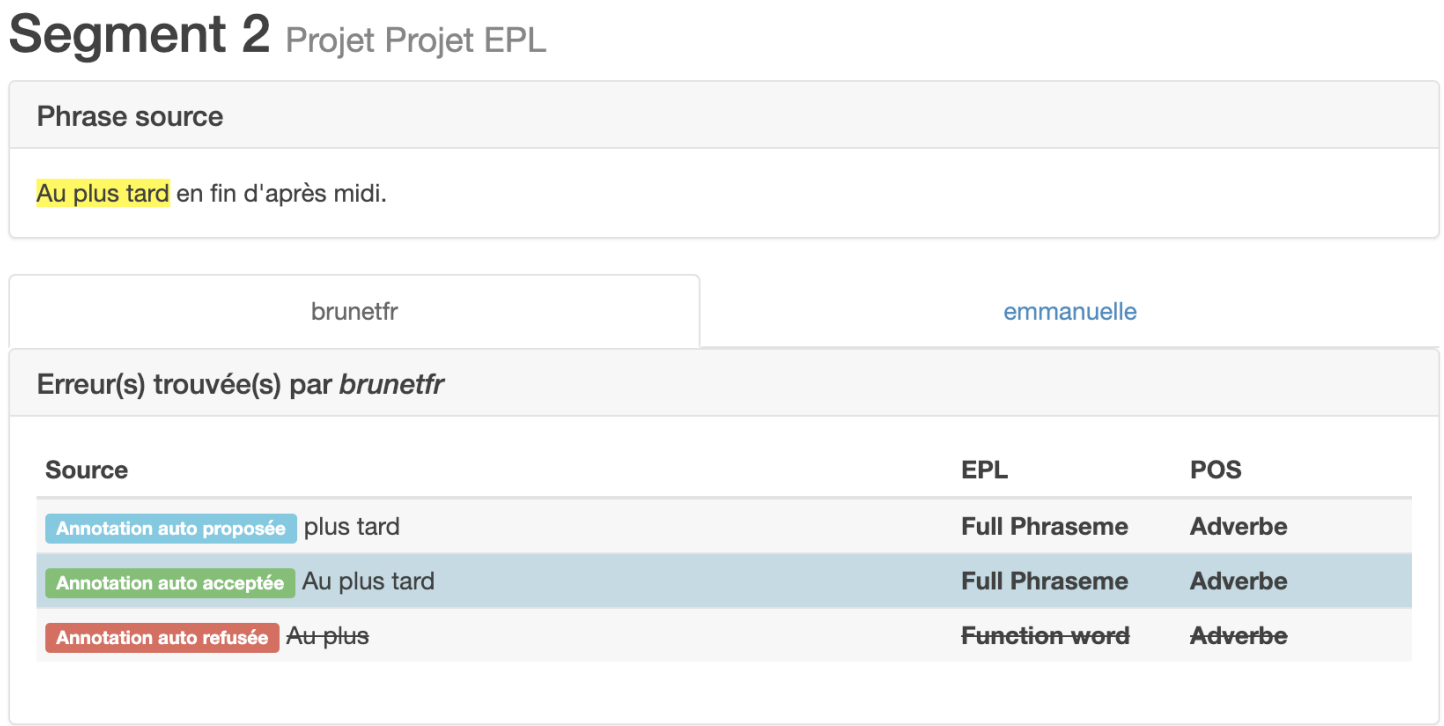

Figure 6 : Affichage des informations pour la Supervision

Enfin, le mode supervision de la plateforme (Figure 6), permet une visualisation des annotations d'EPL en fonction de leur provenance, si l'annotation a été réalisée manuellement ou bien si elle est issue d'une proposition automatique, et de leur statut, si l'annotateur l'a acceptée, modifiée, refusée ou bien si l'annotation a été proposée à l'annotateur mais est en attente d'acceptation, de refus ou de modification.

3 Données disponibles

ACCOLÉ propose 3 typologies d'erreurs, celle de Vilar et al. (2006), deux autres issues de MQM-DQF (Lommel, 2018)) et 1 typologie d'annotation des EPL (Tutin & Esperança, 2017) ainsi que 14 corpus FR-GB, 4 corpus monolingues FR/GB et 7 corpus GB-DE (allant des nouvelles journalistiques, à des documents techniques, des brevets, des extraits du BTEC (Basic Travel Expression Corpus) jusqu'à des documents sur le climat ou des textes médicaux) pour un total de 25 corpus (+66,6% en un an), ayant permis la création de 30 projets (+58%). Ceux-ci correspondent à 9 585 phrases (+40,5 %), 184 786 mots sources (+37,6%), 266 078 mots cibles (+132%), pour 34 558 annotations réalisées par 12 annotateurs natifs soit anglais, allemand ou français (+47%). Ces corpus sont structurés selon les SNODEs (Boitet et al., 1988) et sont disponibles sur demande au format XML ou JSON. Une fonction permet de rechercher dans ces corpus les types d'erreurs. Au moment de la rédaction, nous continuons de travailler sur la recherche de modèle d'erreurs.

4 Conclusion

Dans cet article, nous présentons les fonctionnalités supplémentaires apportées à la plateforme ACCOLÉ. Certains corpus annotés disponibles ont déjà été utilisés pour une comparaison linguistique de la qualité de la traduction de différents systèmes de TA (Esperança-Rodier et Becker, 2018), ainsi que pour l'analyse de la qualité de systèmes de TA simultanée ou après complétion de la phrase entière (Elbayad et al., 2020). ACCOLÉ permet désormais d'étudier les phénomènes de traductions liés aux EPLs grâce à la production de corpus annotés en EPL afin de contribuer à la recherche en linguistique. Un projet d'annotation d'EPL et d'évaluation de TA français-polonais d'EPL vient d'être accepté par NeuroCog/Pôle Grenoble Cognition et débutera en janvier 2021.

ACCOLÉ permet également d'avoir une interface multi-cibles comme fréquemment utilisé lors des grandes campagnes d'évaluation. Outre l'implémentation de la recherche d'erreurs au sein des corpus par patrons morphosyntaxiques pour la partie annotation d'erreurs, nous étudions pour la partie annotation d'EPL, la possibilité de modifier le dictionnaire en cours de traitement : correction des entrées du dictionnaire, ajout d'entrées en fonction des annotations manuelles.

Enfin, nous comptons également intégrer des informations du web sémantique afin de proposer plus de format d'annotations, ou bien de données liées aux annotations.

Références

Boitet, C. et Zaharin, Y. (1988). Representation trees and string- tree correspondences. In *Proceedings of international Conference on Computational Linguistics* COLING-88, 59-64.

Coavoux, M. et Crabbé, B. (2017). Représentation et analyse automatique des discontinuités syntaxiques dans les corpus arborés en constituants du français. *Actes de la 24e conférence sur le Traitement Automatique des Langues Naturelles, Jun 2017, Orléans, France. pp.77-92*

Elbayad, M., Ustaszewski, M., Esperança-Rodier, E., Brunet-Manquat, F., Verbeek, J. et Besacier, L. (2020). Online Versus Offline NMT Quality: An In-depth Analysis on English–German and German–English. *Accepté à COLING 2020*.

Esperança-Rodier, E. et Becker, N. (2018). Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs. Proceedings of the 4th day on « *Traitement Automatique des Langues et Intelligence Artificielle » - TALIA 2018 Day of the plate-forme Intelligence Artificielle (PFIA 2018)*. Nancy, France, 6 juillet 2018. Edited by Didier Schwab et Pierre Zweignebaum.

Esperança-Rodier, E., Brunet-Manquat, F., et Eady, S. ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. *Translating and the computer 41, Nov 2019, Londres, United Kingdom*

Hellmann, S., Lehmann, J., Auer, S., Brümmer, M. Integrating NLP using Linked Data. *2th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*

Lommel, A., et Alan, K. M. (2018). Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). *13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers). Vol. 2.*

Popović, M. (2018) Error Classification and Analysis for Machine Translation Quality Assessment. *Moorkens J., Castilho S., Gaspari F., Doherty S. (eds) Translation Quality Assessment. Machine Translation: Technologies and Applications, vol 1. Springer, Cham*

Tutin, A, Esperança-Rodier, E. (2017) La difficile identification des expressions polylexicales dans les textes : critères de décision et annotation. "*La phraséologie française : débats théoriques et dimensions appliquées (didactique, traduction et traitement informatique)*", *Sep 2017, Arras, France*.

Vilar, D., Xu, J., D'Haro, L.F. and al. (2006) Error analysis of statistical machine translation output. *5th International Conference on Language Resources and Evaluation*. 97-702.

Alignement temporel entre transcriptions et audio de données de langue japhug

Cécile Macaire[1 2]
(1) Langues et Civilisations à Tradition Orale (LACITO), CNRS-Sorbonne Nouvelle, France
(2) Laboratoire d'Informatique de Grenoble (LIG), CNRS-Université Grenoble Alpes, France
cecile.macaire@live.com

RÉSUMÉ

La collection Pangloss héberge un corpus de langue japhug : plus de 30 heures d'enregistrements audio (et dans une moindre mesure vidéo) accompagnés de transcriptions. La grande majorité des transcriptions, réalisées par le linguiste Guillaume Jacques, spécialiste de la langue, ne comportaient pas d'informations concernant l'alignement texte-son : les seuls points de référence étaient le début et la fin des enregistrements, dont la durée va de 22 secondes à 33 minutes. Le présent exposé présente la façon dont des chronocodes (balises indiquant l'alignement texte-son) ont été ajoutés au niveau de la phrase. La chaîne de traitement consiste en une application de l'outil d'alignement forcé MAUS, doublé d'une étape de vérification manuelle. Ce travail, réalisé sans intervention du linguiste, permet désormais la consultation phrase par phrase de la transcription, la citation d'une phrase spécifique par le biais d'une référence DOI, ainsi que l'utilisation des documents pour entraîner un modèle acoustique en vue de la transcription phonémique automatique de cette langue *à faibles ressources*. Cette tâche illustre l'utilité de collaborations entre linguistes et ingénieur·e·s informatiques pour la documentation linguistique.

ABSTRACT

Time alignment between transcriptions and audio of Japhug language data

The Japhug language corpus in the Pangloss Collection consists of a set of over thirty hours of audio recordings (and some video recordings) transcribed in the International Phonetic Alphabet (IPA). The vast majority of the transcriptions, made by linguist Guillaume Jacques, a specialist in the language, did not have detailed information about text-sound alignment : the only points of reference were the beginning and end of each of the recordings, which range from 22 seconds to 33 minutes in length. This presentation provides an account of the way in which detailed temporal information has been added (time codes indicating text-to-sound alignment at the level of the sentence) using the MAUS segmentation tool, and how the temporal information has been added to the transcripts. This task was entirely conducted without hands-on implication by the linguist. The result is that it is now possible (i) to consult texts sentence-by-sentence, (ii) to quote a specific sentence via a DOI, and last but not least, (iii) to use the documents for training an acoustic model for automatic phonemic transcription of untranscribed audio. The alignment task reported here illustrates the usefulness of collaborations between linguists and computer engineers for linguistic documentation.

MOTS-CLÉS : Alignement temporel, japhug, documentation linguistique.

KEYWORDS: Temporal alignment, Japhug, linguistic documentation.

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 9–22,

1 Introduction

1.1 La collection Pangloss

La collection Pangloss (Michailovsky et al., 2014) [1] est une archive multimédia de langues en danger. Elle accueille des ressources linguistiques dans plus de 150 langues : enregistrements audio et vidéos accompagnés, pour certains, de transcriptions et annotations. Certains ont exprimé la crainte de voir les archives orales devenir des «cimetières de données» (“data graveyards” : Gippert et al., 2006, 4, 12-13) ; la collection Pangloss est bien plutôt un «jardin de données», dans lequel les corpus sont enrichis, au fil des ans, d'annotations de plus en plus complètes. Le dépôt de fichiers dans divers formats est accepté, y compris de simples fichiers texte brut ou des manuscrits scannés, mais l'objectif est de parvenir au stade de textes dotés d'un appareil critique complet (traductions, notes...), encodé en XML. Ainsi, le présent exposé relate l'ajout d'un alignement temporel entre transcriptions et audio pour le corpus de langue japhug, et en expose brièvement les enjeux.

1.2 La langue japhug et le corpus japhug

Le japhug est une langue sino-tibétaine parlée par environ 10 000 personnes dans la province du Sichuan (Chine). Le travail de Guillaume Jacques au sujet de cette langue depuis une vingtaine d'années a abouti à de nombreuses publications, dont une thèse (Jacques, 2004), un dictionnaire japhug-chinois-français (Jacques, 2016), et une étude phonético-phonologique (Jacques, 2019). Dans la collection Pangloss, il y a environ 400 enregistrements sonores en langue japhug, dont la durée va de 22 secondes pour le plus court à 33 minutes pour le plus long. La majorité des enregistrements font moins de 10 minutes. Trois locuteurs sont présents. Les enregistrements contiennent des contes, aussi bien que des témoignages au sujet de la culture locale : descriptions de la faune et de la flore, informations sur les techniques agricoles, la construction de maisons, le tissage... (Jacques, 2015). Avant les tâches décrites ici, les transcriptions réalisées par G. Jacques ne possédaient pas d'informations concernant l'alignement texte-son, hormis quelques documents qui possèdent un alignement phrase par phrase (par exemple le récit *Les trois sœurs* [2]). Il était de ce fait difficile d'accéder à l'enregistrement d'une phrase spécifique de la transcription. Un travail a été réalisé afin d'aligner les transcriptions avec les enregistrements audio, et d'ajouter ces informations temporelles (chronocodes) au niveau de la phrase.

2 Méthode

2.1 Outil d'alignement automatique : MAUS

Il existe divers outils pour effectuer un alignement forcé entre un fichier audio et sa transcription : SailAlign (Katsamanis et al., 2011), EasyAlign (Golman, 2011), etc. Le choix s'est porté sur MAUS (*Munich Automatic Segmentation*) (Schiel, 1999; Kisler et al., 2017), qui a été utilisé avec succès dans le contexte de la documentation linguistique (Strunk et al., 2014). MAUS calcule une segmentation

1. https://pangloss.cnrs.fr
2. https://doi.org/10.24397/pangloss-0003357

phonétique et un étiquetage, sur la base du signal audio fourni, et d'une transcription phonologique encodée selon des conventions spécifiques : en Alphabet Phonétique International, ou dans son équivalent ASCII selon les conventions SAMPA. La transcription phonologique doit être fournie dans un fichier BPF au format BAS Partitur (extension .par). MAUS permet de traiter un discours aussi bien lu que spontané d'une langue non répertoriée.

2.2 Chaîne de traitement

Dans le présent travail, il n'a pas été fait usage de fonctionnalités avancées de MAUS, telles que la prise en charge des variantes de prononciation. L'outil a été utilisé de façon relativement élémentaire. La figure 1 explique les étapes nécessaires à l'ajout des balises temporelles au niveau de la phrase dans les fichiers XML.

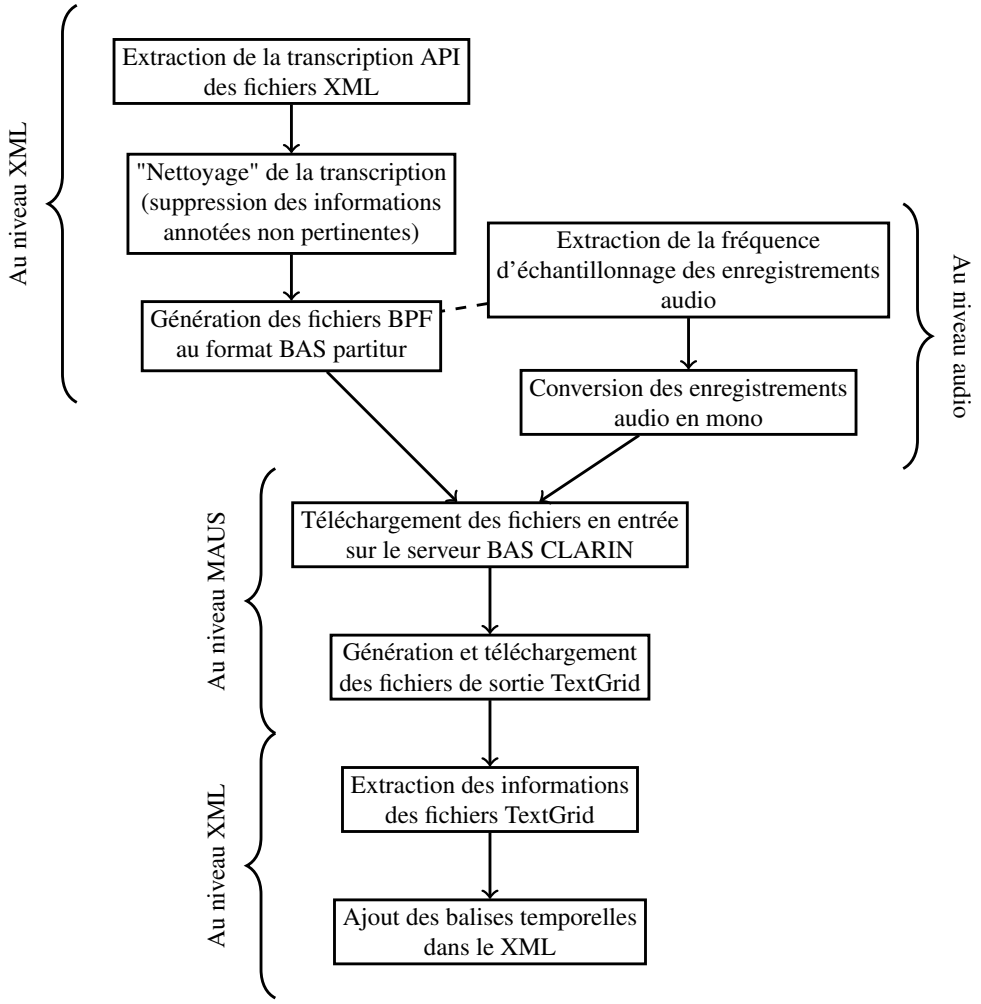

FIGURE 1 – Les différentes étapes réalisées par le script python.

En partant du corpus japhug, la première étape a consisté en l'extraction des données, ici, des transcriptions au format XML (cf. figure 2), et leur prétraitement : un toilettage pour supprimer certaines informations non prises en compte par l'outil et vérifier l'encodage de certains caractères.

Dans le corpus japhug, les mots empruntés au chinois mandarin sont notés en écriture chinoise, ce qui a le double avantage de les identifier avec grande précision et de les faire ressortir visuellement (ainsi que dans l'encodage). On voit ainsi dans la figure 3 le mot 机耕道 'sente pour tracteurs'. Cette pratique est claire et intuitive pour les linguistes spécialistes de langues de Chine, mais comme elle

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE TEXT SYSTEM "https://cocoon.huma-num.fr/schemas/Archive.dtd">
<TEXT id="crdo-JYA_HIST-06-TSHAT-QAZO-KALAG" xml:lang="jya">
    <HEADER>
</HEADER>
    <FORM>
tshɤt qaʑo kɤ-lɤɣ ɲɯ-qha-nɯ ma
ɲɯ-rkaŋ-nɯ qhe aʁɤndɯndɤt ʑo sɯŋgɯ tu-ɕe-nɯ qhe
rdɤstaʁ ra pjɯ-tʂaβ-nɯ qhe
tɯrme tu-xtsɯɣ ɲɯ-ŋu
tʂu ra pjɯ-sɯ-Nqhi ɲɯ-ŋu qhe
jigendao 机耕道 nɯnɯ
mkhɯrlu tʂu nɯra pjɯ-sɯNqhi tɕe ɲɯ-qha-nɯ qhe
tɕe ɯ-kɯ-ftɕa ɲɯ-rkɯn
</FORM>
</TEXT>
```

FIGURE 2 – Transcription de l'histoire 6 *tshAt* [3] dans un fichier XML.

n'est pas explicitée dans la courte documentation disponible au sujet du corpus, elle constitue un obstacle pour les utilisateurs non spécialistes du domaine.

Les transcriptions pré-traitées ont ensuite été converties au format XML de la collection Pangloss : le découpage en phrases, indiqué à l'origine par des retours chariot, se trouve encodé (dans la logique du *texte structuré logiquement*) sous forme d'éléments <S> (*Sentence* : niveau de la phrase), puis alignées avec l'enregistrement audio. Enfin, ces informations temporelles récupérées ont été ajoutées dans les documents XML, sous forme de chronocodes au niveau de la phrase. La figure 3 montre le début du récit *tshAt* [3] obtenu après ce traitement. On y reconnaît les trois premières phrases du texte, cette fois-ci explicitement codées comme telles (éléments <S>) et chacune dotée d'une balise <AUDIO> contenant le début et la fin de l'intervalle audio correspondant. On relèvera également l'ajout d'une balise <TITLE> (on s'est contenté, pour l'heure, de recopier l'identifiant *a minima* que constituait le nom du fichier fourni par le linguiste) et d'une balise <SOUNDFILE> indiquant le nom du fichier audio correspondant.

```
<?xml version="1.0" ?>
<!DOCTYPE TEXT SYSTEM "https://cocoon.huma-num.fr/schemas/Archive.dtd">
<TEXT id="crdo-JYA_HIST-06-TSHAT-QAZO-KALAG" xml:lang="jya">
        <HEADER>
                <TITLE>crdo-JYA_HIST-06-TSHAT-QAZO-KALAG</TITLE>
                <SOUNDFILE href="hist-06-tshAt-qaZo-kAlAG.wav"/>
        </HEADER>
        <S id="S001">
                <AUDIO start="0.21" end="2.48"/>
                <FORM kindOf="phono">tshɤt qaʑo kɤ-lɤɣ ɲɯ-qha-nɯ ma</FORM>
        </S>
        <S id="S002">
                <AUDIO start="2.82" end="5.47"/>
                <FORM kindOf="phono">ɲɯ-rkaŋ-nɯ qhe aʁɤndɯndɤt ʑo sɯŋgɯ tu-ɕe-nɯ qhe </FORM>
        </S>
        <S id="S003">
                <AUDIO start="5.8" end="7.52"/>
                <FORM kindOf="phono">rdɤstaʁ ra pjɯ-tʂaβ-nɯ qhe</FORM>
```

FIGURE 3 – Début de la transcription de l'histoire 6 *tshAt* [3] annotée avec les balises temporelles phrase par phrase.

L'ensemble de ces étapes a été automatisé par un script python : **xml_info_japhug.py** (Macaire,

3. https://doi.org/10.24397/pangloss-0003420

2020) [4]. Il faut tout d'abord extraire les transcriptions. La fonction **extract_information(xml_file)** extrait les informations de la balise $< FORM >$. La transcription est une chaîne de caractères bruts, les phrases étant séparées par un saut de ligne. La chaîne de caractères est découpée par ligne, et nettoyée.

La langue Japhug n'est pas répertoriée dans MAUS, il a donc fallu utiliser le paramètre "Language independent" et son inventaire de caractères acceptés. Spécifiquement au corpus Japhug, de nombreux caractères spéciaux posaient problème car ils ne sont pas pris en charge dans l'inventaire utilisé. La solution a été de remplacer les caractères inconnus par MAUS par ceux répertoriés.

La notation de la langue japhug comporte des marques accentuelles, qui font référence à des événements prosodiques. Les formes accentuées, ó, ú, á, í, etc. sont remplacées par o, u, a, i. De plus, certains mots possèdent un préfixe et/ou suffixe séparé de la base par un '-'. Pour l'alignement phonémique, ces particules ont été "raccrochées" à la base pour ne former qu'un ensemble : [jɤ-ari-a] devient [jɤaria].

2.3 Informations écartées lors du traitement

Certaines informations ajoutées par le linguiste ne font pas partie du discours mais sont interprétées comme telles par MAUS. Celles-ci posent problèmes lors de la génération des fichiers BPF au format BAS partitur :

- Indications temporelles posées dans le fil du document : par exemple *(201)* correspondant à 2 minutes 01.
- Division de la transcription en parties (paragraphes) au moyen de suites de '=', 'x', '-', '*'.
- Caractères chinois : utilisation d'une librairie python hanzidentifier pour les identifier.
- Signes de ponctuations (' ?', ' !', '.', etc.)
- Autres informations : '(calque)', '(faute)', '(ideoph)', ' ???', 'X', etc. regroupés dans une liste *irrelevant_annotations* (cf. figure 9).

Lorsque le fichier est débarrassé d'informations autres que la transcription brute du discours, on peut créer les fichiers BPF au format BAS partitur.

2.4 Génération des fichiers BPF au format BAS partitur

La fonction **create_par(text, file, frame_rate)** (cf. figure 13) génère les fichiers BPF, fichiers en entrée pour MAUS. Le texte est découpé en mots, et chaque mot se voit attribuer un identifiant unique. La fréquence d'échantillonnage du fichier audio correspondant à la transcription est récupérée en parallèle grâce à la fonction **get_sampling_rate(file)** (cf. figure 11), et les fichiers son stéréo sont convertis en mono au moyen de la librairie python pydub, appelée par la fonction **convert_mono(path, path_export, file)** (cf. figure 12). Les fichiers sont ensuite téléchargés en entrée dans MAUS. Une requête HTTP télécharge les fichiers vers le serveur, effectue la segmentation et retourne un fichier TextGrid en sortie. Des paramètres sont à spécifier en entrée, comme montré en figure 4.

Dans le script python, cela correspond à la ligne de commande :

4. https://gitlab.com/macairec/alignement_texte_son_japhug

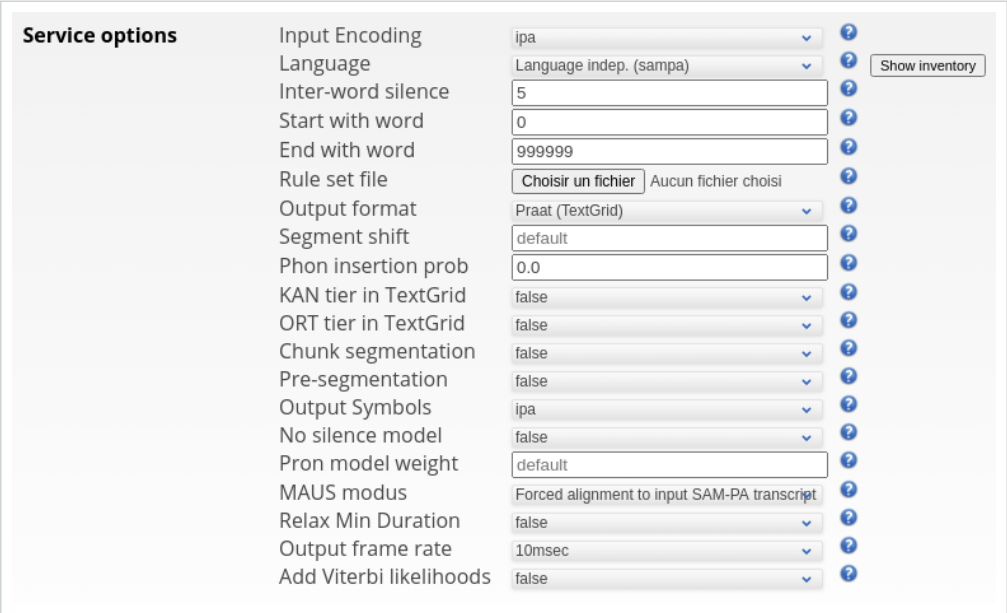

FIGURE 4 – Options en entrée de MAUS.

```
'curl -v -X POST -H \'content-type: multipart/form-data\' -F SIGNAL=@' \
+ path_wav + wav_file + ' -F LANGUAGE=sampa -F INSKANTEXTGRID=false -F ' \
'MODUS=align -F RELAXMINDUR=false -F OUTFORMAT=TextGrid -F ' \
'TARGETRATE=100000 -F ENDWORD=999999 -F STARTWORD=0 -F INSYMBOL=ipa -F ' \
'PRESEG=false -F USETRN=false -F BPF=@' + path_par + par_file + \
' -F MAUSSHIFT=10 -F INSPROB=0.0 -F INSORTTEXTGRID=false -F MINPAUSLEN=5 -F ' \
'OUTSYMBOL=ipa -F WEIGHT=default -F NOINITIALFINALSILENCE=false -F ' \
'ADDSEGPROB=false ' \
'\'https://clarin.phonetik.uni-muenchen.de/BASWebServices/services/runMAUS\''
```

FIGURE 5 – Ligne de commande dans le script python.

2.5 Extraction des chronocodes des TextGrid et ajout dans le fichier XML

Le but de ce script est d'ajouter des informations temporelles pour chaque ligne de transcription. Chaque ligne est identifiée grâce à un saut de ligne matérialisé par **\n**. La fonction **extract_info_textGrid(textgrid)** (cf. figure 14) crée un dictionnaire qui comprend deux éléments : un identifiant pour chaque phrase de transcription, et les balises temporelles de début et de fin d'apparition dans l'audio (exprimées en secondes). En effet, la fonction récupère le paramètre **xmin** du premier mot de la phrase, et le paramètre **xmax** du dernier mot de la phrase, reconnaissable par **\n**.

La figure 6 ci-dessous est un exemple d'un fichier TextGrid : ici la phrase commence par le mot [stu] à 0.44 seconde de l'audio, et se termine par le mot [babɯ] à 1.73 secondes.

La dernière étape consiste à ajouter les informations temporelles de début et de fin de phrase dans le fichier XML comprenant la transcription. La fonction **add_xml_info(timecode, wav_file, xml_file)** (cf. figure 15) récupère le dictionnaire précédemment créé, puis la transcription du fichier XML. A chaque ligne dans le fichier XML, on associe un identifiant, les balises temporelles, et la transcription. Certains caractères tels que les séparateurs de parties de transcriptions ('-', '=', '*') sont répertoriés dans des balises **NOTE**. La figure 7 illustre le fichier en sortie créé. Le script permet

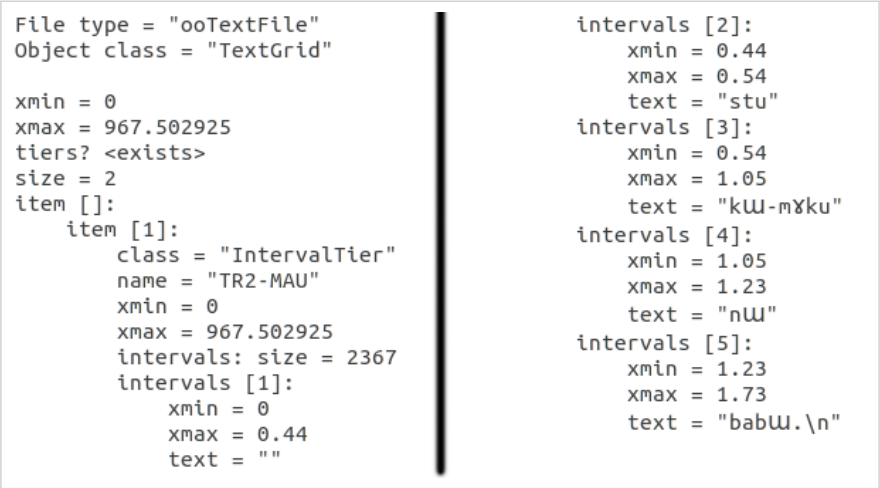

FIGURE 6 – Exemple d'un fichier TextGrid généré par MAUS (Schiel, 1999), (Kisler et al., 2017).

```
<?xml version="1.0" ?>
<!DOCTYPE TEXT SYSTEM "https://cocoon.huma-num.fr/schemas/Archive.dtd">
<TEXT id="crdo-JYA_HIST140425_KWFCI" xml:lang="jya">
        <HEADER>
                <TITLE>crdo-JYA_HIST140425_KWFCI</TITLE>
                <SOUNDFILE href="hist140425_kWfCi.wav"/>
        </HEADER>
        <S id="S001">
                <AUDIO start="0.53" end="4.23"/>
                <FORM kindOf="phono">kɯɕɯŋgɯ tɕe, kɯfɕi nɯnɯra wuma ʑo, nɤkinɯ,</FORM>
        </S>
        <S id="S002">
                <AUDIO start="4.23" end="7.0"/>
                <FORM kindOf="phono">nɯ kɯ-fse mɤ-kɯ-taŋ tu-nɤma-nɯ pjɤ-ŋgrɤl tɕe,</FORM>
        </S>
```

FIGURE 7 – Exemple d'un début de fichier XML avec les balises temporelles.

également d'ajouter deux informations dans l'en-tête : le titre du fichier et la référence du fichier audio correspondant. La seconde figure 8 ci-dessous montre l'utilisation d'une balise **NOTE** qui encapsule des informations portées par le linguiste mais qui n'appartiennent pas à proprement parler à la forme d'une phrase.

```
<NOTE message="================" xml  :lang="fr"/>
<S id="S022">
        <AUDIO start="75.3" end="76.61"/>
        <FORM kindOf="phono">tɕe mɤʑɯ phuɲi tu.</FORM>
</S>
<S id="S023">
        <AUDIO start="77.38" end="79.92"/>
        <FORM kindOf="phono">phuɲi nɯ li si ŋu tɕeri, mɤ-mbro.</FORM>
</S>
<S id="S024">
        <AUDIO start="80.26" end="80.99"/>
        <FORM kindOf="phono">tɕhi tɤ-mbro,  </FORM>
</S>
<S id="S025">
        <AUDIO start="80.99" end="83.07"/>
        <FORM kindOf="phono">tɯrme tɯ-fsu ɕaŋtaʁ tu-mbro mɤ-cha.</FORM>
</S>
```

FIGURE 8 – Partie d'un fichier XML annoté comprenant une balise NOTE.

3 Problèmes rencontrés

Quelques difficultés ont été rencontrées lors de la création du script et de son application. Tout d'abord, afin d'exclure les caractères non pertinents lors de la création des fichiers BPF, il a fallu regarder un à un chaque fichier qui posait problème lors du lancement de leur segmentation. Le principal défi a été de travailler sur un ensemble documentaire sans connaissance de la langue japhug, ni du chinois, et sans supervision préalable. Le corpus regroupant des ressources s'étalant sur une vingtaine d'années, il était difficile de comprendre d'emblée chaque information annotée. La présence d'un inventaire comprenant l'ensemble des normes et des informations supplémentaires annotées par le linguiste aurait permis un gain de temps considérable dans la construction du script. Un tel inventaire serait particulièrement utile pour les personnes faisant du TAL (Traitement Automatique des Langues). Les conventions varient d'un fichier à l'autre. La mise en place de normes d'écriture est un point essentiel, et permettrait une uniformisation du corpus.

3.1 Décalages entre annotation et audio

De plus, la transcription pouvait parfois présenter des disparités avec l'audio. Par exemple, certains mots dits n'ont pas été transcrits. La majeure partie du temps, cela correspondait aux mots [nɤkinɯ] et [tɕendɤre], plus communément appelés "gap-fillers". Ces choix effectués par le linguiste éloignent l'annotation du signal audio. De plus, certaines parties d'audio n'étaient pas transcrites, ou encore certaines transcriptions comprenaient une partie absente de l'audio (cf. fichier crdo-JYA_HIST-06-HUCHEMENTS1.xml par exemple).

3.2 Bruits parasites

Un point essentiel concerne les fichiers audio. Certains d'entre eux comportent des sons parasites, que MAUS considèrent comme de la parole. L'alignement résultant en est faussé. Il s'agit par exemple d'hésitations (« hum », « heu », etc.), de toussotements et éternuements, de chuchotements en arrière plan, de paroles d'une personne tierce dont le discours n'est pas retranscrit, de bruits de rue, ou de bruits parasites très légers en début et fin de fichier. Sur l'ensemble du corpus, près de la moitié des enregistrements audio présentaient de tels sons. La solution la plus efficace a été de "mettre en silence" les parties parasites. La fonction **mute_sound(path_wav, wav_file)** effectue cette opération ; il suffit de spécifier les intervalles de temps en secondes à traiter.

4 Conclusion et perspectives

Par l'utilisation d'un outil de segmentation automatique, MAUS, les informations temporelles pertinentes ont été extraites et ajoutées dans les transcriptions du corpus japhug. Les documents enrichis sont d'ores et déjà consultables via l'interface web de la collection Pangloss.

La méthodologie développée dans ce projet comporte plusieurs limitations, notamment car certains éléments des enregistrements audio (bruits parasites) et des transcriptions (annotations non adaptées) faussaient l'alignement produit, rendant nécessaire une intervention manuelle. Sur la totalité du corpus

(environ 400 fichiers), seuls 6 fichiers n'ont pas pu être corrigés, du fait que la transcription n'était pas complète.

Il n'a pas été mené d'évaluation objective de la qualité des résultats obtenus : cela aurait nécessité un point de comparaison (par exemple, un alignement temporel de référence, qui aurait été établi de façon entièrement manuelle, pour les mêmes documents). Néanmoins, de l'avis du linguiste déposant (Guillaume Jacques), les résultats obtenus au terme du processus sont concluants pour les données de langue japhug : aucune erreur d'alignement n'a été détectée.

4.1 Intérêt pour la consultation des données

L'existence de chronocodes au niveau de la phrase facilite une meilleure navigation au sein du corpus. En effet, tous les documents de la collection Pangloss ont été dotés d'identifiants DOI (Digital Object Identifier). A ce sujet, on consultera un article (Vasile et al., 2020) qui revient sur les problématiques d'identification de la ressource, et sur le contexte «socio-scientifique» actuel, pour mettre en perspective le choix d'attribuer un DOI à chaque document de la Collection Pangloss. Il est possible d'ajouter au DOI un numéro de phrase, après un dièse : par exemple, `https://doi.org/10.24397/pangloss-0003357#S10` amène droit à la dixième phrase (unité S) du document en question. Cela permet par exemple de citer un exemple précis dans une publication scientifique : les lecteurs peuvent accéder en un clic à l'exemple, dans son contexte d'origine. L'existence d'un chronocode au niveau de la phrase permet d'écouter le passage concerné, sans devoir naviguer tant bien que mal au sein d'un fichier audio dont la durée, rappelons-le, peut dépasser 30 minutes.

4.2 Intérêt pour l'utilisation du corpus japhug dans des expériences en Traitement Automatique des Langues Naturelles

Les transcriptions enrichies de chronocodes ont aussitôt été utilisées dans des expériences de reconnaissance automatique de la parole : transcription automatique de langues à faibles ressources. Ces expériences, qui s'inscrivent dans le fil de travaux exploratoires réalisés depuis plusieurs années (Adams et al., 2017, 2018; Michaud et al., 2018, 2019; Wisniewski et al., 2020a; Michaud et al., 2020), font l'objet d'un exposé lors des présentes Journées scientifiques (Wisniewski et al., 2020b).

Remerciements

Vifs remerciements à Séverine Guillaume, Laurent Besacier et Alexis Michaud pour avoir guidé le travail décrit ici, et au Groupement de Recherche CNRS LIFT (Linguistique Informatique, Formelle et de Terrain), qui m'a orienté vers ce travail dans le cadre d'une mis en relation entre linguistes et étudiants de Traitement Automatique des Langues.

Ce travail a reçu un soutien financier dans le cadre du projet « La documentation computationnelle des langues à l'horizon 2025 » (ANR-19-CE38-0015-04).

Références

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365, Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648.

Adams, O., Cohn, T., Neubig, G., and Michaud, A. (2017). Phonemic transcription of low-resource tonal languages. In *Proceedings of the 2017 Australasian Language Technology Association Workshop (ALTA 2017)*, pages 53–60, Brisbane, Australia. https://halshs.archives-ouvertes.fr/halshs-01656683.

Gippert, J., Himmelmann, N., and Mosel, U. (2006). Language documentation : What is it and what is it good for. In *Essentials of language documentation*, volume 178, pages 1–30. Walter de Gruyter, Berlin.

Golman, J.-P. (2011). EasyAlign : An automatic phonetic alignment tool under Praat. In *Proceedings of InterSpeech 2011*, Florence. International Speech Communication Association.

Jacques, G. (10/05/2015). "Mon travail sur le japhug (1)," Panchronica,. https://panchr.hypotheses.org/264(ISSN2494-775X).

Jacques, G. (2004). *Phonologie et morphologie du japhug (rGyalrong)*. PhD thesis, Univ. Paris-Diderot/Paris VII Paris.

Jacques, G. (2016). Dictionnaire Japhug-chinois-français Version 1.1. https://halshs.archives-ouvertes.fr/halshs-01003734.

Jacques, G. (2019). Japhug. *Journal of the International Phonetic Association*, 49(3) :427–450.

Katsamanis, A., Black, M., Georgiou, P., Goldstein, L., and Narayanan, S. (2011). SailAlign : Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang.*, 45(C) :326–347.

Macaire, C. (2020). Script python. https://gitlab.com/macairec/stage_lacito.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages : the Pangloss Collection. *Language Documentation and Conservation*, 8 :119–135. https://halshs.archives-ouvertes.fr/halshs-01003734.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow : experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12 :393–429.

Michaud, A., Adams, O., Cox, C., and Guillaume, S. (2019). Phonetic lessons from automatic phonemic transcription : preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*, Melbourne. https://halshs.archives-ouvertes.fr/halshs-02059313.

Michaud, A., Adams, O., Guillaume, S., and Wisniewski, G. (2020). Analyse d'erreurs de transcriptions phonémiques automatiques d'une langue « rare » : le na (mosuo). In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole, Traitement Automatique des Langues*

Naturelles, Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 451–462, Nancy, France. ATALA. https://hal.archives-ouvertes.fr/hal-02798572.

Schiel, F. (1999). Automatic phonetic transcription of nonprompted speech. In *Proc. Int. Cong. Phon. Sci*, pages 607–610.

Strunk, J., Schiel, F., Seifart, F., et al. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In *LREC*, pages 3940–3947.

Vasile, A., Guillaume, S., Aouini, M., and Michaud, A. (2020). Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger. *I2D - Information, données & documents*, 2 :156–175.

Wisniewski, G., Guillaume, S., and Michaud, A. (2020a). Phonemic transcription of low-resource languages : To what extent can preprocessing be automated ? In Beermann, D., Besacier, L., Sakti, S., and Soria, C., editors, *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, pages 306–315, Marseille, France. European Language Resources Association (ELRA). https://halshs.archives-ouvertes.fr/hal-02513914.

Wisniewski, G., Michaud, A., Galliot, B., Besacier, L., Guillaume, S., Aplonova, K., and Jacques, G. (2020b). Ouvrir aux linguistes « de terrain » un accès à la transcription automatique. In *Actes des Journées scientifiques du Groupement de Recherche "Linguistique informatique, formelle et de terrain" (LIFT)*, Paris.

5 Annexe : script Python

```
['!)', '"', '"/???/', '(', '(!!!', '(...)', '(?)', '(??)', '(???)', '(agreement',
 '(anticausative)', '(antipassif)', '(associated', '(au', '(autoben)', '(autoben+anticaus)',
 '(autobenefactive)', '(benefactif)', '(calque', '(calque)', '(calque?)', '(cataphoric)',
 '(causatif)', '(causation)', '(causative)', '(causative,', '(cleft)', '(comme)',
 '(complementation', '(complementation)', '(conatif)', '(conative)', '(converb)', '(dative)',
 '(diminutive)', '(en', '(erreur', '(faute)6:40', '(faux:', '(floating)', '(focalisation,',
 '(focus)', '(generic)', '(hybrid?)', '(ideoph', '(ideoph)', '(ideoph.pause).', '(il', '(ils',
 '(imp?', '(imperative', '(in', '(indefinite)', '(infinitive)', '(intensif', '(inverse',
 '(inverse', '(inverse', '(inverse)', '(inverse+associated', '(inverse?', '(nominal', '(pas',
 '(passive)', '(permansive)', '(pluriel)', '(possessive)', '(redp)', '(reduplication',
 '(relation', '(relative)', '(revoir)', '(right', '(superlative)', '(trill)', ')', '*', ',',
 '-', '..', '...', '...,', '....', '/', '/.../', '//', '//?//', '/?/', '/???/', '050)',
 '0805)', '1015)', '200)', '2:00', '?', '?"', '??', '??', '???', '???,', '????', 'X', '\\b',
 'adjectives)', 'causee)', 'ce', 'dans', '(050', 'de', 'de', 'deux', 'deux', 'deux', 'disloc)',
 'emph)', 'entre', 'et', 'fact?', 'fur', 'hybrid)', 'inanimés).', 'inanimés).', 'inverse)',
 "l'original,", 'la', 'lui', 'main)', 'mange', 'manner)(testimonial)', 'mesure)', 'mismatch)',
 'modifiers,', 'motion)', 'motion,', 'n\'avaient', 'particulier)', 'parties', 'passif',
 'preventive)', 'quasi-clivée)', 'qui', 'redp)', 'relative,', 'rien', 'permansif?)', 'sous',
 'strategy)', 'that)', 'tombe', 'tout', 'traduction', 'à', '||||||', '(IMM)', 'khWjNga2',
 '(faute)', '||||||\"', '(demonstrative,', 'pronouns,', 'present)', '(kɯ?', '730)',
 'possessor)']
```

FIGURE 9 – Liste irrelevant_annotations.

```
def extract_information(xml_file):
    """Extract the transcription from xml file"""
    tree = et.parse(xml_file)
    root = tree.getroot()  # lexical resources
    transcript = root.find('FORM').text  # get content from tag FORM
    return transcript
```

FIGURE 10 – Fonction extract_information(xml_file).

```
def get_sampling_rate(file):
    """Get the sample rate of a wav file"""
    with wav.open(file, "rb") as wave_file:
        frame_rate = str(wave_file.getframerate())
    return frame_rate
```

FIGURE 11 – Fonction sampling_rate(file).

```
def convert_mono(path, path_export, file):
    """Convert wav stereo files into mono wav files"""
    sound = AudioSegment.from_wav(path + file)
    sound = sound.set_channels(1)
    sound.export(path_export + file, format="wav")
```

FIGURE 12 – Fonction convert_mono(path, path_export, file).

```
def create_par(text, file, frame_rate):
    """Create PAR file from the transcription previously extracted from xml file"""
    words_tr2 = [e + '\n' for e in text.split('\n') if e]  # split text by lines
    words_tr2 = [i.replace('\ufeff', '') for i in words_tr2]  # remove space at the beginning
    words_tr2 = [i.replace('>', '') for i in words_tr2]  # remove '>'
    digits = [re.findall(r'\(\d+.*\)', l) for l in words_tr2]  # find all the digits of the form '(201)'
    for k, p in enumerate(words_tr2):
        if len(digits[k]) != 0:
            words_tr2[k] = words_tr2[k].replace(digits[k][0], '')  # and delete them
    words_tr2 = [words_tr2[i].split() for i, j in enumerate(words_tr2)]  # split sentences into words
    words_tr2 = [[i for i in nested if (
            i not in irrelevant_annotations and not hanzidentifier.has_chinese(
        i) and 'xxxx' not in i and '---' not in i and '\\c' not in i and 'FICHIER' not in i and '=' not in i)] for
                 nested in
                 words_tr2]  # delete characters for MAUS
    words_tr2 = list(filter(None, words_tr2))  # remove blank lines
    words = [x for z in words_tr2 for x in z if x]  # extract all the words from sublist
    replace_characters = {'-': '', 'g': 'ɡ', 'ʁ': 'R', 'ɯ̽': 'ɯ', 'ɤ̞': 'ɤ', 'ó': 'o', 'ú': 'u', 'G': 'ɣ', '2': 'ø',
                          '4': 'ɾ', 'M': 'ɯ', 'T': 't', 'ʉ': 'y', 'û': 'u', 'é': 'e', 'B': 'b', 'à': 'a', 'á': 'a',
                          ', ': '', 'A': 'a', 'U': 'u', 'ɦ̃': 'h', '@': '', ' (': '(', ') ': ')', '6': 'ɐ', '9': 'œ',
                          '7': 'ɤ', '0': '', 'Z': 'z', 'ů': 'u', 'í': 'i', '1': 'ɨ', '3': ''}
    for k, v in replace_characters.items():
        words = [w.replace(k, v) for w in words]
    words = [w.translate(str.maketrans('', '', s.punctuation)) for w in words]  # delete punctuation
    phonemes = [i for i in words]  # separate each 'letter' of each word
    if len(words) != 0:
        with open(file, 'w+') as f1:  # create the par file
            f1.write('LHD: Partitur 1.3.1' + '\n' + 'SAM: ' + frame_rate + '\n' + 'LBD:' + '\n')
            for i, j in enumerate(words):
                char = ''
                for el in phonemes[i]:
                    char += el + ' '
                f1.write('KAN: ' + str(i) + ' ' + char + '\n')  # add KAN level
            for i, j in enumerate(words):
                f1.write('ORT: ' + str(i) + ' ' + j + '\n')  # add ORT level
            compt = 0
            for i, j in enumerate(words_tr2):  # add TR2 level with '\n' for end of sentence
                for k, m in enumerate(words_tr2[i]):
                    if k == len(words_tr2[i]) - 1:
                        m = m + "\\n"
                        f1.write('TR2: ' + str(compt) + ' ' + m + '\n')
                        compt += 1
                    else:
                        f1.write('TR2: ' + str(compt) + ' ' + m + '\n')
                        compt += 1
        return True
    else:
        return False
```

FIGURE 13 – Fonction create_par(text, file, frame_rate).

```
def extract_info_textGrid(textgrid):
    """Extract information from textGrid generate by MAUS"""
    data = {}
    grid = tg.TextGrid(textgrid)  # extract TextGrid
    time = []
    compt = 0
    for item in grid['TR2-MAU']: # extract timecodes
        if item.text != "" and '\\n' in item.text and len(time) == 0:
            time.append(item.xmin)
            time.append(item.xmax)
        else:
            if item.text != "" and '\\n' not in item.text and len(time) == 0:
                time.append(item.xmin)
            elif '\\n' in item.text and '.' in item.text:
                time.append(round(item.xmax, 2))
            elif '\\n' in item.text:
                time.append(item.xmax)
        if len(time) == 2:
            data[compt] = time
            time = []
            compt += 1
    return data
```

FIGURE 14 – Fonction extract_info_textGrid(textgrid).

```
def add_xml_info(timecode, wav_file, xml_file):
    """Add information for each transcription lines in xml file"""
    tree = etree.parse(xml_file)
    root = tree.getroot()  # lexical resources
    header = root.find('HEADER')
    time_begin = 0
    if header.find("TITLE") is None:  # add information into HEADER
        title = etree.SubElement(header, 'TITLE')
        title.text = xml_file.split('/')[7][:-4]
        sound_file = etree.SubElement(header, 'SOUNDFILE', href=wav_file)
    identifiant = 0
    has_s = False
    sentences = []
    if root.find('FORM') is not None:  # extract transcription from FORM
        transcript = root.find('FORM')
        trans = list(filter(None, transcript.text.split('\n')))
        identifiant = 0
    else:
        has_s = True
        trans = []
        sentences = root.findall('S')
        for i, j in enumerate(sentences):
            if j.find('FORM') is not None:
                identifiant = int(j.attrib['id'][1:])
                time_begin = float(j.find('AUDIO').attrib['end'])
            else:
                trans.append(j.text)
        sentences = [el for el in sentences if el.find('FORM') is None]
    for key, value in timecode.items():
        value[0] += time_begin
        value[1] += time_begin
    num = 0
    for i, j in enumerate(trans):  # for each line
        if '==' in j or 'xxxx' in j or '---' in j or '\\c' in j or '\\b' in j or (
                j.startswith('(') and j.isdigit() and j.endswith(')')) or 'khWjNga2' in j:
            nsmap = {"lang": "fr"}
            tag = etree.SubElement(root, "NOTE", nsmap=nsmap, message=j)  # add NOTE element for specific characters
        elif num < len(timecode):  # add timecode until the end
            if 0 <= identifiant < 9:
                if has_s:
                    tag = sentences[i]
                    tag.set('id', 'S00' + str(identifiant + 1))
                    tag.text = None
                else:
                    tag = etree.SubElement(root, "S", id='S00' + str(identifiant + 1))
             elif 9 <= identifiant < 99:
                 if has_s:
                     tag = sentences[i]
                     tag.set('id', 'S0' + str(identifiant + 1))
                     tag.text = None
                 else:
                     tag = etree.SubElement(root, "S", id='S0' + str(identifiant + 1))
             elif 99 <= identifiant < 999:
                 if has_s:
                     tag = sentences[i]
                     tag.set('id', 'S' + str(identifiant + 1))
                     tag.text = None
                 else:
                     tag = etree.SubElement(root, "S", id='S' + str(identifiant + 1))
             time = etree.SubElement(tag, 'AUDIO', start=str(round(timecode[num][0], 2)),
                                     end=str(round(timecode[num][1], 2)))
             form = etree.SubElement(tag, 'FORM', kindOf="phono")
             form.text = j
             num += 1
             identifiant += 1
         else:
             if not has_s:
                 tag = etree.SubElement(root, "S")
                 tag.text = j
    form = root.find('FORM')
    if form is not None:
        root.remove(form)
    tree.write(xml_file, encoding='utf-8', xml_declaration=True)  # create xml file with the new annotations
    dom = xml.dom.minidom.parse(xml_file)  # or xml.dom.minidom.parseString(xml_string)
    pretty_xml_as_string = dom.toprettyxml()  # correctly print xml file
    pretty_xml_as_string = '\n'.join([line for line in pretty_xml_as_string.split('\n') if line.strip()])
    lines = pretty_xml_as_string.split('\n')
    del lines[1:3]
    lines.insert(1, '<!DOCTYPE TEXT SYSTEM "https://cocoon.huma-num.fr/schemas/Archive.dtd">')
    for k, m in enumerate(lines):
        if '<AUDIO' in m:
            a = m.split(' ')
            if 'end' in a[1]:
                a[1:3] = a[1:3][::-1]
                a[1] = a[1][:-2]
                a[2] = a[2] + '/>'
                lines[k] = ' '.join(a)
    pretty_xml_as_string = '\n'.join(lines)
    pretty_xml_as_string = pretty_xml_as_string.replace("&quot;", "\"")
    with open(xml_file, "wb") as f:
        f.write(pretty_xml_as_string.encode('utf-8'))
```

FIGURE 15 – Fonction add_xml_info(timecode, wav_file, xml_file).

Classification des catégories grammaticales sur deux corpus longitudinaux d'enfants

Andrea Briglia[1, 2] Jérémi Savage[1] Giovanni Pirrotta[2] Massimo Mucciardi[2]
(1) Université Paul Valéry, Montpellier, France
(2) Université de Messine, Messine, Italie
prenom.nom@univ-montp3.fr, prenom.nom@unime.it

RÉSUMÉ

Cet article analyse deux suivis longitudinaux de deux enfants du projet CoLaJE: une annotation automatique des parties du discours a été appliquée à chaque énoncé (15'000 en total) en adoptant le standard des « Universal Dependencies » comme référence et « stanza », un librairie Python, comme outil d'analyse. L'âge et le taux d'erreur ont servi comme base pour la création de neuf strata: réduire la dimension du corpus nous permet de rendre intérpretables les groupements crées avec une méthode non-supervisée, EM clustering. Regrouper en clusters les énoncés des enfants annotés en parties du discours aide à mieux cibler le développement des catégories grammaticales au cours du temps: deux exemples concernants le développement de la cohérence morphosintaxique sont proposés, ainsi que deux exemples concernants l'évolution de la relation entre l'usage de pronoms et des noms. Une discussion nale des résultats et des limites de cette recherche est ensuite proposée.

ABSTRACT

Classi cation of grammatical categories in two longitudinal chlidren corpora. This article analyses two child spoken language longitudinal corpora from the CoLaJE project: a parts of speech automatic annotation was applied to each sentence (15'000 in total) using « Universal Dependencies » as a standard of reference and "stanza", a Python library, as an analysis tool. Age and error rate were used as criteria for the creation of nine strata: reducing the size of the corpus helps to make more easily interpretable clusters created with EM, an unsupervised method. Aim of the article is to propose a way to target the development of grammatical categories over time: two examples concerning the development of morphosintactic coherence are proposed, as well as two examples concerning the evolution of the relationship between the use of pronouns and nouns. A final discussion of the preliminary results and limitations of this research is then proposed.

MOTS-CLÉS : corpus d'enfants ; acquisition du français L1 ; développement syntaxique ; clusterisation EM

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 23–31,

KEYWORDS: child spoken language corpora ; first language acquisition ; syntax development ; EM clustering

1 Objectifs et hypothèse de recherche

Le projet ANR « CoLaJE » (Morgenstern & Parisse, 2012) consiste en sept corpora d'enfants francophones filmés une heure par mois, tous les mois, dès l'âge d'un an jusqu'à environ 5 ans. L'ensemble de données est disponible en libre accès et fait partie de la branche française de CHILDES[1] . Nous avons choisi cette base de données parce que – à ce jour – elle est la plus complète sur les plans qualitatif et quantitatif. Par ailleurs, nous estimons que l'échantillonnage effectué dans la collecte mensuelle des données est conforme aux indications de fiabilité énoncées par Tomasello et Stahl (2004). Chaque corpus a été codé en CHAT et transcrit en pho (ce que l'enfant prononce) et – pour certains corpora dont les deux qu'on utilise en cette étude – mod (ce que l'enfant aurait dû prononcer selon la norme de prononciation standard), ce qui nous permet d'uniformiser les données phonético/phonologiques, de les contextualiser pour mieux les interpréter et, enfin, de pouvoir y appliquer des traitements automatiques.

Dans la présente contribution, nous nous focalisons sur les corpora d'« Adrien » et « Madeleine » car ils sont les plus complets : nous avons extrait chaque ligne en format .csv, ensuite nous avons choisi de commencer par la transcription no 8 (1an 11mois; 14jours) pour Adrien et la no 3[2] (1 ;01 ;10) pour Madeleine, puisque pour les précédentes il était difficile de distinguer entre les mots et les simples suites de syllabes correspondantes à l'étape du « babillage canonique » et du « babillage diversifié » dans le développement de la production de la parole du jeune enfant (Sauvage, 2015). Nous avons au total 26 enregistrements et 8214 énoncés pour Adrien et 25 enregistrements et 7168 énoncés pour Madeleine. Nous avons choisi le « Universal Dependencies » (de Marneffe et al., 2006, 2008, 2014) comme modèle de référence d'analyse du langage en parties du discours, principalement parce que nous avions déjà eu recours à ce modèle (Briglia et al., 2020). Ce choix nous a conduit à adopter "stanza", un outil d'analyse du langage majoritairement entraîné en utilisant les UD. « stanza » est une des bibliothèques de TAL disponible en langage Python, développée par l'Université de Stanford: puisque le système d'annotation automatique ne reconnaît pas les caractères spéciaux de l'API (Alphabet Phonétique International), nous l'avons appliqué sur le tiers CHI (transcription orthographique): ce choix implique une forte confiance envers l'interprétation des transcripteurs : il est néanmoins possible de consulter – énoncé par énoncé – toutes les différences entre CHI – pho –mod. La qualité de l'annotation produite par « stanza » est élevée et, pour la plupart des tâches, son score est meilleur que celui de ses concurrents (e.g UDPipe, spaCy), comme le montre le tableau numéro 2 « Neural pipeline performance comparisons on the Universal Dependencies (v2.5) test treebanks" (Qi et al., 2020).

Puisque le langage de l'enfant se caractérise par une forte variabilité et reste imprévisible à court et moyen termes et puisque UD et « stanza » ont été conçus pour le langage des adultes, il nous a

[1] https://childes.talkbank.org/access/French/

[2] https://ct3.ortolang.fr/data/colaje/madeleine/; https://ct3.ortolang.fr/data/colaje/adrien/

semblé nécessaire d'opérer un contrôle manuel de quatre-vingts énoncés pour chaque enfant (ce qui représente environ 1% du total) équitablement répartis au fil du temps, afin de comprendre l'effective fiabilité de l'outil pour cette application. Nous avons remarqué que certaines répétitions du même mot - typiques lorsque l'enfant cherche de cibler l'apprentissage d'un mot donné – étaient parfois codés comme NOUN – ADJ, alors qu'il s'agissait soit de deux noms communs, soit de deux adjectives (par exemple, l'énoncé « des grands grands arbres » était codé comme « DET-ADJ-NOUN-NOUN »), « ouais » était codé comme SYM et, de façon plus générale, les nombreuses exclamations des enfants (par exemple « ah ! », « oh ») ou des surnoms comme « papi » ou « mémé » étaient codés différemment selon leur contexte. Mais ces exceptions sont tout à fait faciles à comprendre dans leur contexte et, en tout cas, ne représentent qu'un faible pourcentage du total des productions. En fait, pour effectuer une analyse syntaxique pertinente sur un corpus longitudinal d'enfant il est indispensable de comprendre avant tout si ce que l'enfant dit – que ce soit au niveau phonético/phonologique ou syntaxique - est conforme ou non aux normes linguistiques des adultes ou pas. C'est pourquoi nous nous sommes appuyés sur une précédente étude où le SPVR (Sentence Phonetic Variation Rate) avait été calculé (Briglia A. et al., 2020). Le taux obtenu sera le résultat d'une comparaison entre les tiers « pho » et « mod » : pour ce faire, nous avons mis en place un algorithme indiquant si le premier est équivalent au deuxième ou non, en donnant comme réponse 0 ou 1 et la distance de Levenshtein relative.

1.1 Corpus recueilli et méthodologie d'analyse

Notre but est de fournir un outil d'évaluation du développement de la syntaxe basé sur des associations et des distributions. Afin de savoir comment les catégories syntaxiques évoluent pendant le temps et le taux d'erreur de ces dernières, nous divisons l'ensemble des données en 9 strata (LL, LM, LH, ML, MM, MH, HL, HM, HH) selon trois classes temporelles successives (représentées par la première lettre) et trois classes d'erreur (représentée par la deuxième lettre). Par exemple, LL veut dire que le strata représente la première tranche d'âge et le taux d'erreur le plus bas (c'est-à-dire Low < 33.3%).

code	STRATA	TIME (age)	SPVR
1	LL	1.01 - 2.09	<=33%
2	LM	1.01 - 2.09	>33% and <=66%
3	LH	1.01 - 2.09	>66%
4	ML	2.10 - 2.61	<=33%
5	MM	2.10 - 2.61	>33% and <=66%
6	MH	2.10 - 2.61	>66%
7	HL	2.70 - 3.53	<=33%
8	HM	2.70 - 3.53	>33% and <=66%
9	HH	2.70 - 3.53	>66%

Figure 1: Madeleine

code	STRATA	TIME (age)	SPVR
1	LL	1.97 - 2.64	<=33%
2	LM	1.97 - 2.64	>33% and <=66%
3	LH	1.97 - 2.64	>66%
4	ML	2.71 - 3.39	<=33%
5	MM	2.71 - 3.39	>33% and <=66%
6	MH	2.71 - 3.39	>66%
7	HL	3.46 - 4.33	<=33%
8	HM	3.46 - 4.33	>33% and <=66%
9	HH	3.46 - 4.33	>66%

Figure 2: Adrien

Ainsi, nous voyons que la 1 ère tranche d'âge représente le parcours de Madeleine à partir de l'âge d'un an jusqu'à deux ans et 1 mois. Cette tranche compte 12 enregistrements, pour un total de 1956 énoncés analysés. La 2 ème tranche d'âge est constituée de 6 enregistrements (c'est-à-dire six mois consécutifs) et 2765 énoncés. Enfin, la 3 ème et dernière tranche d'âge est composée de 7 enregistrements et 2447 énoncés. La première tranche compte deux fois plus d'enregistrements que les suivantes parce que l'enfant à cet âge parle très peu et la quasi-totalité de ce qu'il dit sont des mots isolés (il s'agit du stade « holophrastique » de son développement syntaxique, où un mot comme « eau » peut vouloir signifier un énoncé entier comme « je veux boire »). Les six premiers enregistrements ne comptent que 66 énoncés, alors que le 7 ème enregistrement (1 an et six mois) en compte 187. En effet, à partir de 18 mois et jusqu'à 30 mois on assiste à une période d'explosion lexicale où l'enfant va apprendre plusieurs mots par jour. En particulier, de 18 mois à 24 mois l'enfant développe un lexique de plus en plus riche mais qui n'arrive pas encore à combiner au niveau syntaxique (on parle de langage « télégraphique », d'une syntaxe construite autour d'un mot pivot, par exemple dans les énoncés à deux mots), alors qu'à partir de 24 mois jusqu'à 36 mois on assiste au développement de la grammaire, avec une phase dite d'« explosion grammaticale » à partir de 30 mois environ (Sekali M., 2012).

Pour la première tranche d'âge d'Adrien, on compte 1709 énoncés repartis en 9 enregistrements, pour la 2 ème tranche d'âge on compte 2623 énoncés en 9 enregistrements et la 3 ème tranche compte 3882 énoncés en 8 enregistrements. Nous obtenons donc un total de 25 enregistrements et 7168 énoncés pour Madeleine, et 27 enregistrements et 8214 énoncés pour Adrien. C'est en cherchant un équilibre entre les étapes du développement et les enregistrements disponibles que nous avons choisi de diviser en trois parties le nombre total d'enregistrement des deux enfants pour mieux organiser l'analyse. On peut remarquer que les enregistrements d'Adrien commencent à presque deux ans et que les transcriptions en tier « mod » se terminent pour Madeleine à l'âge de 3;5 ans alors que pour Adrien elles continuent jusqu'à l'âge de 4;4 ans : nous avons ainsi choisi de décaler le début d'analyse pour Adrien, ce qui nous a permis d'obtenir des tranches temporelles plus proches et, en conséquence, plus comparables.

Il reste une différence remarquable entre le développement langagier des deux enfants : celui d'Adrien pourrait être considéré comme normé, alors que celui de Madeleine est sans doute plus

rapide que la moyenne (Morgenstern & Parisse, 2012). De façon plus générale, il est couramment accepté que les filles parlent souvent mieux que les garçons à un même âge, que ce soit au niveau qualité ou quantité de parole. Mais ces derniers rattrapent cet écart entre 4 et 6 ans. Pour résumer, nous montrerons de façon simplifiée l'évolution du nombre des mots par énoncé et l'évolution du nombre de mots différents par énoncé ci-dessous (en bleu Madeleine et en rouge Adrien). Ensuite, nous discuterons si l'analyse en clusters conduite avec EM peut être mise en relation avec ces graphes. Il faut se rappeler que dans ces graphes les trois points temporels sur les abscisses représentent un décalage de presque un an: malgré elle soit plus jeune, Madeleine montre un meilleur développement, donc si les deux lignes auraient été temporellement alignées, l'écart aurait été majeur (ce qui est conforme aux résultats de l'étude principale sur ce corpus par Morgenstern & Parisse, 2012).

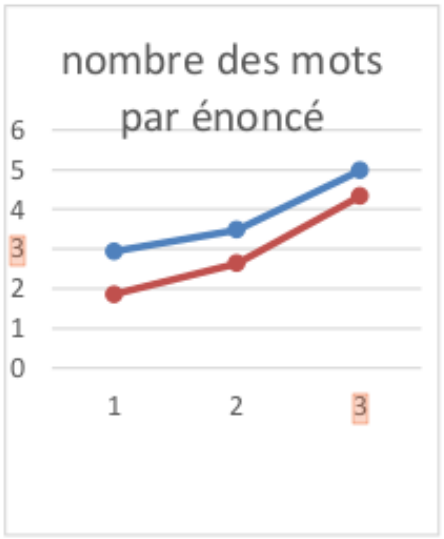

Figure 3

Figure 4

1.2 Traitement automatique

Nous calculons le F (Fisher) et le p-value relatif pour chaque partie du discours (POS tag) en supposant que la distribution des occurrences discrètes sous-jacente est de type Poisson (nous avions d'abord essayé avec une distribution de type Gauss mais sans résultat convaincant). Ensuite le nombre d'énoncés est divisé en plusieurs groupes grâce à la méthode non-supervisée EM (Expectation Maximization). Nous avons choisi cette méthode parmi d'autres parce qu'elle s'adapte bien, selon nous, à la quantité et à la typologie de nos données. Son fonctionnement est le suivant : le nombre optimal de groupes est obtenu en suivant les deux étapes d'un algorithme itératif qui termine son calcul lorsque l'ajout ultérieur d'un groupe aboutit à une amélioration négligeable dans la fonction de vraisemblance logarithmique (le seuil a été fixé à 5% de la vraisemblance à n-1). En sélectionnant les parties du discours plus significatives pour les regroupements, EM est influencé en

amont par une distinction fondamentale entre les classes ouvertes (ADJ, ADV, INTJ, NOUN, PROPN, VERB) et les classes fermées (ADP, AUX, CCONJ, DET, NUM, PART, PRON, SCONJ) : les premières contiennent beaucoup plus d'éléments que les deuxièmes, pour lesquelles le nombre est fixé. Pourtant, la fréquence entre les deux classes ne varie pas autant, ce qui est confirmé dans le développement de la parole (voir les valeurs des variables mean associées à chauqe POS tag dans les tableaux ci-dessous). Toute étude concernant la distribution des fréquences des mots et leurs interactions devrait en fait garder à l'esprit l'importance de l'équilibre entre syntaxe et sémantique comme principe organisateur du langage (Zipf, 1949 ; Lestrade, 2017).

1.3 Résultats préliminaires

La cohérence morphosyntaxique est plus élevée dans les clusters en HL, HM par rapport à ceux dans les couches L et M, ce qui est conforme aux résultats d'une étude précédente (Parisse et al., 2010). On peut remarquer que les parties du discours PRON, VERB, SCONJ – qui pourraient être considérées comme marqueurs des phrases plus longues – augmentent leur importance (voir la variable mean) au fil du temps, il est également à noter que les groupes qui reconnaissent des énoncés syntaxiquement proches font aussi partie des classes d'erreurs et d'âge différentes, par exemple, dan le cas d'Adrien:

2452 escargot tout chaud

ɛskaʁgo tu ʃo

didago to so

En MH

6746 une souris verte

yn suʁi vɛʁtə

yn ʦoji vatə

En HH On peut ensuite noter comme le NOUN et le PROPN – qui indiquent une personne ou une chose concrète et sont souvent utilisés dans le même contexte de la parole adulte (répétition du langage adressé à l'enfant) – sont au fur et à mesure remplacés par PRON (une catégorie grammaticale plus abstraite pour indiquer choses et personnes) en termes du nombre relatif d'occurrence (réf aux tableaux 1 et 2). L'âge à laquelle ce changement devient visible est environ 3 ans et correspond à l' âge individuée dans des études précédentes sur le développement de la capacité de produire des énoncés en forme transitive à partir des énoncés en forme intransitive (Childers &Tomasello, 2001). Par exemple, dans le cas d'Adrien,

1973 sait pas faire maman	sɛ pa fɛʁ mamɑ̃	te ta pa mamɑ̃
2915 est caché papa	ɛ kaʃe papa	e kaɕe papa
3674 il peut pas ouvrir la porte	i pø pa uvʁiʁ la pɔʁt	i pu pa uvij a pɔt

En MH , on peut remarquer le manque du pronom « je », alors que à l'âge 3_09_09 dans cet énoncé les pronoms sont bien utilisés

5339 moi j' aime plutôt celle-ci maman elle aime plutôt celle-ci

mwa ʒɛm plyto sɛlsi mamɑ̃ ɛl ɛm plyto sɛlsi

mwa zem pyto sɛsi mamɑ̃ ɛl øm pytosɛti

En observant la valeur de la variable « mean », on peut remarquer comme la valeur de PRON augmente au fil du temps en dépit de la valeur de NOUN : chez Madeleine ce développement est plus rapide que chez Adrien.

Madeleine présente un développement plus ordonné puisque dans les mêmes trois strata considérés pour Adrien, EM arrive dans ce cas à identifier la structure syntaxique la plus basique : sujet (ou pronom) – verbe. Les énoncés de Madeleine présentent constamment plus VERB and AUX que celles d'Adrien, sauf pour le troisième strata : ce qui colle avec les graphes montrés en Figure 3 et 4, où nous pouvons remarquer que les énoncés de la petite fille sont à la fois plus longs et lexicalement plus riches.

Par exemple, en HL, énoncé 186, cluster 2

euh ça je sais pas comment l'ouvrir

ø sa ʒə sɛ pa komɑ̃ luvʁiʁ

œ sa ʒə sɛ pa komɑ̃ luvʁiʁ

On peut remarquer comment le dernier strata HL de Adrien ressemble à celui de Madeleine : on en pourrait déduire que lorsque les quatre premières parties du discours classées comme plus significatives par EM contiennent NOUN, PRON, VERB et AUX alors l'enfant devrait présenter des énoncés syntaxiquement conformes à la norme adulte.

STRATA LL			STRATA ML			STRATA HL		
POS_tags	Mean	# sentences	POS_tags	Mean	# sentences	POS_tags	Mean	# sentences
INTJ	0,126	611	CCONJ	0,048	851	PRON	1,157	1762
DET	0,095	611	PRON	0,133	851	DET	0,321	1762
ADP	0,013	611	NOUN	0,220	851	VERB	0,788	1762
NOUN	0,468	611	AUX	0,052	851	NOUN	0,419	1762
SYM	0,023	611	VERB	0,157	851	SCONJ	0,149	1762
ADV	0,563	611	NUM	0,035	851	ADP	0,230	1762
PROPN	0,016	611	SYM	0,020	851	AUX	0,208	1762
PRON	0,025	611	ADV	0,828	851	ADV	0,727	1762
VERB	0,023	611	DET	0,086	851	ADJ	0,091	1762
X	0,020	611	PROPN	0,029	851	CCONJ	0,120	1762
CCONJ	0,023	611	ADP	0,034	851	SYM	0,022	1762
SCONJ	0,011	611	X	0,026	851	NUM	0,085	1762
AUX	0,007	611	INTJ	0,176	851	X	0,018	1762
NUM	0,103	611	ADJ	0,012	851	PROPN	0,033	1762
ADJ	0,000	611	SCONJ	0,011	851	INTJ	0,162	1762

Tableau 1: Adrien

STRATA LL			STRATA ML			STRATA HL		
POS_tags	Mean	# sentences	POS_tags	Mean	# sentences	POS_tags	Mean	# sentences
NOUN	0,73	707	X	0,12	1452	NOUN	0,68	1171
VERB	0,38	707	NOUN	0,61	1452	PRON	0,95	1171
PRON	0,27	707	DET	0,45	1452	DET	0,53	1171
DET	0,28	707	VERB	0,70	1452	X	0,14	1171
ADP	0,16	707	PRON	0,81	1452	VERB	0,85	1171
X	0,07	707	AUX	0,20	1452	ADP	0,41	1171
AUX	0,04	707	ADV	0,60	1452	AUX	0,20	1171
ADV	0,34	707	ADP	0,30	1452	SCONJ	0,13	1171
NUM	0,03	707	SCONJ	0,07	1452	CCONJ	0,16	1171
SYM	0,02	707	ADJ	0,10	1452	ADV	0,55	1171
SCONJ	0,01	707	CCONJ	0,10	1452	PROPN	0,05	1171
ADJ	0,04	707	NUM	0,09	1452	NUM	0,12	1171
PROPN	0,04	707	PROPN	0,05	1452	INTJ	0,16	1171
CCONJ	0,02	707	SYM	0,01	1452	SYM	0,02	1171
INTJ	0,09	707	INTJ	0,08	1452	ADJ	0,11	1171

Tableau 2: Madeleine

Pour conclure, le regroupement en clusters pourrait être une façon d'améliorer la compréhension du développement de la syntaxe en proposant une meilleure visualisation de comment une partie du discours évolue dans le temps.

La prochaine étape de ce travail sera d'appliquer cette étude aux autres enfants du projet CoLaJE dans le but de vérifier si les généralisations proposées dans cet article puissent être confirmées pour le développement des autres enfants.

Références

Briglia A., Mucciardi M., Sauvage J. (2020). « Identifying the speech code through statistics: a data-driven approach ».

Childers J., Tomasello M. (2001). « The Role of Pronouns in Young Children's Acquisition of the English Transitive Construction » Developmental Psychology, Vol. 37. No. 6, 739-748.

Dempster A.P., Laird N.M., Rubin D.B. (1977). « Maximum likelihood from incomplete data via the EM algorithm ». Journal of the Royal Statistical Society. Series B: Methodological 39: 1-38.

Lestrade S. (2017). « Unzipping Zipf's law ». PLoS ONE 12(8).

MacWhinney B. (2000). « The CHILDES project: Tools for analyzing talk. 3rd edition ». Mahwah, NJ: Lawrence Erlbaum Associates

Morgenstern A., Parisse C. (2012). « The Paris corpus ». Journal of French language studies /Volume 22/ Special Issue. 7-12. Cambridge University Press (https://www.ortolang.fr/market/corpora/colaje)

Morgenstern A., Sekali M. « What can child language tell us about prepositions? ». Jordan Zlatev, Marlene Johansson

Falck, Carita Lundmark and Mats Andr en. Studies in Language and Cognition, Cambridge Scholars Publishing, 261-275

Parisse C., Le Normand M. T. (2000) « How children build their morphosyntax: The case of French ». Journal of Child Language, Cambridge University Press (CUP), 27, 267-292.

Qi P., Zhang Y., Bolton J., Manning C. D. (2020). « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages ». Association for Computational Linguistics (ACL) System Demonstrations.

Sauvage J. (2015). L'acquisition du langage. Un système complexe. Louvain-la-Neuve : Academia.

Sekali M. (2012). « First language acquisition of French grammar (from 10 months to 4 years old) ». French Language Studies 22, 1-6.

Tomasello M., Stahl, D. (2004). « Sampling children's spontaneous speech: How much is enough? » Journal of Child Language, 31:101–121.

Création d'un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines

Pablo Ruiz Fabo Delphine Bernhard Carole Werner
Université de Strasbourg, LiLPa UR 1339, 67000 Strasbourg, France
{ruizfabo,dbernhard,wernerc@unistra.fr}

RÉSUMÉ

Nous présentons des travaux en cours vers la création d'un corpus diachronique de pièces de théâtre en alsacien pour la période 1870-1940, publiquement disponible, encodé selon les recommandations de la Text Encoding Initiative (TEI) et suivant les principes FAIR pour la création de données de la recherche. Le corpus sera utile aux recherches en sociolinguistique historique et analyse dramatique. Nous décrivons le travail effectué en vue des pratiques FAIR et introduisons des questions de recherche en modélisation TEI de variables pertinentes pour l'analyse linguistique et dramatique. De façon générale, la création du corpus est un exemple des difficultés du travail avec les langues peu dotées. Particulièrement, le corpus présente de l'alternance codique et d'énormes défis pour l'identification automatique des variantes orthographiques, sur lesquels nous aimerions échanger avec la communauté.

ABSTRACT

Creating a FAIR corpus of Alsatian theater and orthographic normalization of non-contemporary varieties

We present work in progress towards creating a diachronic corpus of theater plays in Alsatian. The corpus is publicly available under an open license, encoded according to the Text Encoding Initiative (TEI) guidelines and strives to follow FAIR principles for scholarly data development. We describe our work towards FAIR practices and introduce research questions on the TEI modeling of variables relevant for sociolinguistic and drama analysis. This corpus creation effort exemplifies difficulties related to working with low-resource languages. The corpus shows code-switching and huge challenges for the automatic identification of orthographical variants, which we would like to discuss with the community.

MOTS-CLÉS : corpus, variation, alternance codique, langues peu dotées, TEI, théâtre alsacien.

KEYWORDS: corpus, variation, code-switching, under-resourced languages, TEI, Alsatian theater.

1 Introduction

Le projet MeThAL, « Vers une macroanalyse du théâtre en alsacien [1] », est en train de créer un corpus encodé en TEI (TEI Consortium, 2020) de pièces de théâtre en alsacien pour la période 1870-1940 [2] ;

1. Site du projet : https://methal.pages.unistra.fr/

2. Entre 1871 et 1918, l'Alsace est politiquement rattachée à l'Empire allemand. Le besoin d'auto-détermination des Alsaciens « par rapport au reste du monde allemand » (Huck *et al.*, 2007, 12) passera notamment par le théâtre alsacien et la mise en scène et création de l'Alsace. La date-borne supérieure correspond à l'annexion de l'Alsace au III[e] Reich.

la pièce fondatrice du théâtre dialectal en alsacien, le *Pfingstmontag* de J. G. Arnold (1816), fait également partie du corpus du fait de son importance et son influence dans les pièces plus récentes. Un volume de 50 pièces ou 400 000 tokens est visé. Le corpus est public [3] et suit des principes FAIR ou *Findable*, *Accessible*, *Interoperable*, *Reusable* (Wilkinson *et al.*, 2016). Dans la mesure où le corpus permet de documenter les pratiques langagières de son époque, il aidera à examiner des questions de sociolinguistique historique de l'Alsace (cf. Huck *et al.*, 2007; Huck, 2015). L'encodage permettra une analyse des types de personnages et de la variation linguistique telle que représentée dans leurs paroles selon leur âge, sexe, statut social ou origine, et facilitera aussi l'étude d'aspects formels de la technique dramatique.

Nous présentons des travaux en cours sur la modélisation des données et sur l'identification de variantes orthographiques, nécessaire à cause de l'énorme variabilité dans la représentation écrite de l'alsacien. Des questionnements se posent concernant la création de données linguistiques ouvertes, l'encodage de ressources multilingues qui présentent de l'alternance codique et les méthodes de traitement des langues peu dotées, notamment sur l'identification de variantes orthographiques dans un contexte de ressources linguistiques limitées.

L'article est structuré comme suit : La section 2 présente notre procédure d'encodage TEI et démarche FAIR et nos questionnements autour de la modélisation de variables sociales décrivant les personnages. La section 3 décrit le degré de variation orthographique présent dans le corpus (3.1) ainsi que des cas d'alternance codique (3.2). La section 4 aborde la question de l'identification automatique des variantes dans ce type de corpus.

2 Modélisation et FAIRisation des données

Cette section décrit nos sources, notre procédure d'encodage TEI et nos efforts d'adoption des principes FAIR. La modélisation des descripteurs socio-économiques des personnages est ensuite abordée, ainsi que des possibilités d'encodage TEI de la variation orthographique et de l'alternance codique.

2.1 Sources du corpus

La source principale du corpus est une collection représentative d'environ 150 pièces en alsacien numérisées en 2019 par la Bibliothèque nationale et universitaire (Bnu) à Strasbourg [4]. C'est une ressource électronique fondamentale mais qui demande des améliorations afin de faciliter la recherche linguistique et littéraire : les pièces sont disponibles comme des fichiers d'image, sans balisage, et sans OCR pour la plupart. Nous avons sélectionné un sous-ensemble des pièces visant la variété d'époques et de sous-genres dramatiques [5] et nous avons commencé son océrisation et encodage TEI.

3. Le corpus est mis à jour graduellement sur `https://git.unistra.fr/methal/methal-sources`
4. Voir `https://numistral.fr/fr/theatre-alsacien` (lien *[Découvrir]* pour explorer la collection)
5. Le rendu sur *Drama Corpora* de nos pièces encodées en donne un aperçu : `https://dracor.org/als`

2.2 Procédure d'encodage TEI

Le standard TEI permet la modélisation d'éléments d'analyse dramatique ainsi que de phénomènes linguistiques comme la variation et l'alternance codique. Après océrisation et validation manuelle du texte reconnu, notre encodage TEI s'effectue par une transformation automatique d'une sortie hOCR [6] de Tesseract [7]. Des indices typographiques et de mise en page dans cette sortie reflètent les divisions en acte et scène, répliques et didascalies. Le format est plus variable pour les listes de personnages et les pages de titre, qui fournissent des renseignements essentiels pour les analyses sociolinguistiques et thématiques, ainsi que pour les métadonnées bibliographiques. Afin de gérer ces contenus, nous les avons transcrits manuellement dans une base de données. Nos scripts d'encodage fusionnent ces informations avec la sortie hOCR pour créer les versions TEI. La figure 1 présente la chaîne de traitement.

Notre automatisation de l'encodage TEI repose sur des règles de transformation créées manuellement. Nous voudrions à l'avenir évaluer l'applicabilité de méthodes d'apprentissage automatique, en nous inspirant des travaux de Khemakhem *et al.* (2017, 2018) pour l'encodage TEI de dictionnaires avec des CRF (champs aléatoires conditionnels), qui exploitent la typographie et la mise en page pour la prédiction de la structure TEI. Il serait pertinent de comparer la productivité permise par une telle approche et par notre chaîne de traitement actuelle.

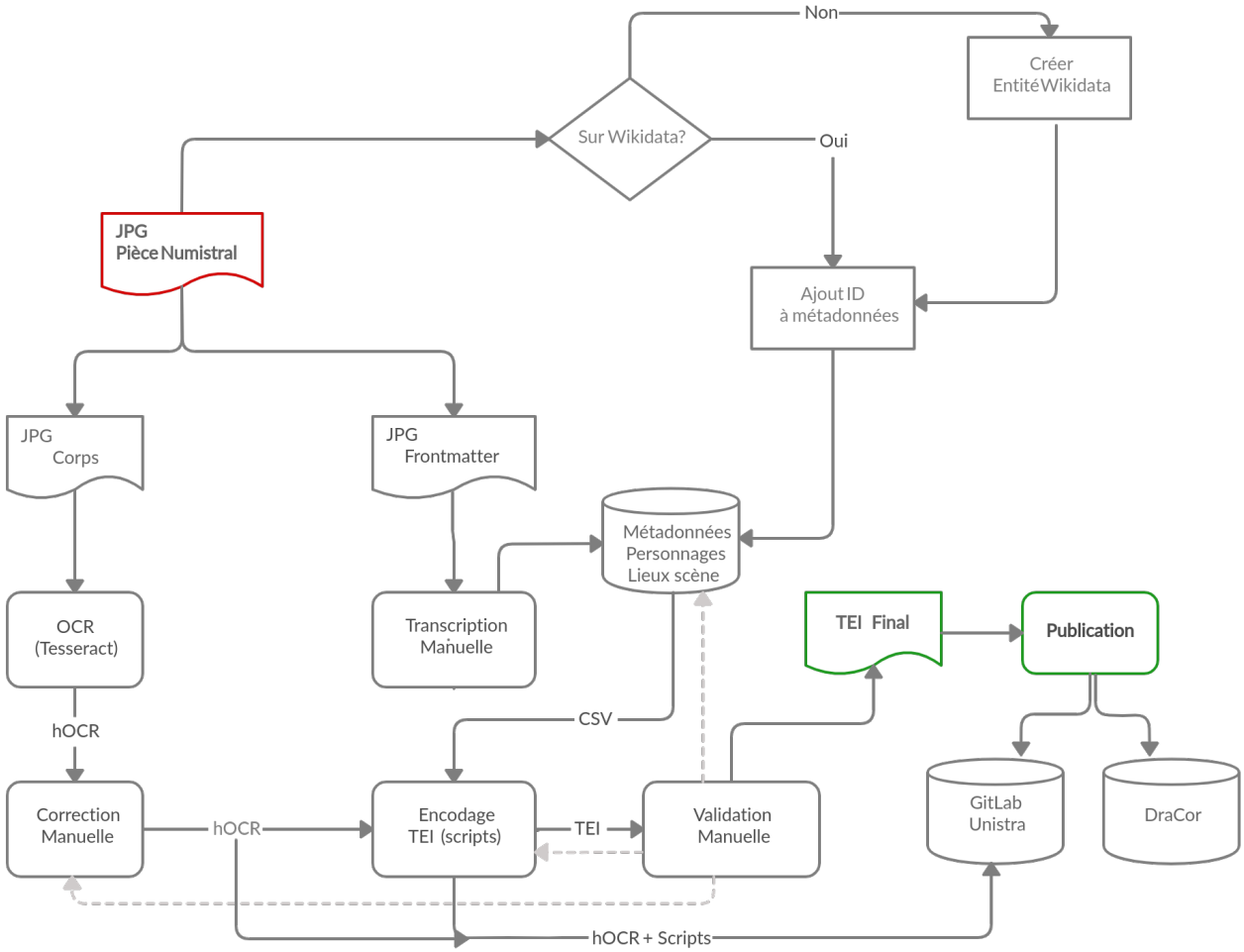

FIGURE 1 – Chaîne de traitement

6. Pour le standard hOCR, voir `http://kba.cloud/hocr-spec/1.2/`
7. `https://github.com/tesseract-ocr/tesseract`

2.3 FAIRisation

Nous visons la création d'un corpus FAIR. Nous avons travaillé sur son interopérabilité et réutilisabilité, et entrepris de premiers pas vers la trouvabilité et accessibilité. Ont contribué à l'interopérabilité l'adoption du standard TEI et l'utilisation d'identifieurs Wikidata pour les pièces et les auteurs, incluant notre création des nouvelles entités Wikidata nécessaires [8]. Concernant la réutilisabilité, chaque pièce est publiée sous une licence ouverte. Pour promouvoir la transparence du processus de prétraitement et d'encodage, les scripts et ressources créés pour traiter chaque pièce, ainsi qu'un wiki pour documenter nos pratiques, sont publiés sur nos dépôts git [9].

Le corpus a des métadonnées riches, en accord avec les requis FAIR pour la trouvabilité des ressources (Wilkinson *et al.*, 2016, 4). Or, il manque à ce jour des identifieurs persistants (DOI ou semblables). Deux options seront considérées dans ce sens : le service d'exposition de données Nakala (Huma-Num, 2020) et le service TEI2Zenodo (Wagner, 2020). C'est aussi par le biais d'une plate-forme d'exposition de données que le corpus sera rendu conforme au critère FAIR d'accessibilité, qui met l'emphase sur l'accès aux données et métadonnées par des programmes informatiques, avec des protocoles de communication standard. Une accessibilité dans un sens moins technique est déjà garantie car le corpus est disponible sur GitLab et sur la plate-forme DraCor[5] (Fischer et Börner, 2019). Celle-ci permet, profitant du balisage TEI, l'accès programmatique à des éléments structurels des pièces (p. ex. toutes les répliques par des femmes ou toutes les didascalies) via une API HTTP.

2.4 Encodage de la variation orthographique et de l'alternance codique

Le corpus doit permettre la comparaison du contenu des pièces, de sorte à faciliter l'analyse de tendances dans les sujets abordés selon diverses variables. À cette fin, la variation orthographique des pièces (voir section 3 pour des exemples) doit être neutralisée ; l'identification automatique des variantes d'un même lexème [10] est un vrai défi, discuté dans la section 4.

Une fois le lexème identifié, la TEI propose des façons naturelles d'encoder la relation entre la variante et son lexème. Une option serait de créer des identifieurs uniques pour les lexèmes du corpus et les donner dans un attribut `@xml:id`. Une autre option serait d'effectuer une normalisation des variantes vers un norme concrète et d'utiliser un élément `<choice>` dont les fils `<orig>` et `<reg>` contiendront la variante originale et normalisée respectivement.

Concernant l'alternance codique, un encodage de base consiste à créer des éléments `<seg>` avec un attribut `@xml:lang` pour le code de la langue de la séquence ; nous avons déjà utilisé cette option dans l'encodage de *D'r Poetisch Oscar* par Marie Hart :

```
<sp who="#oscar">
  <speaker>OSCAR:</speaker>
  <p>Un Sie han m'r wieder d'rzue verholfe, Madame Lewermann,
  Sie ellein verstehn min poetisch Empfinde.
  <seg xml:lang="fre">Vous êtes ma muse</seg>.</p>
</sp>
```

8. Nous avons travaillé sur un sous-ensemble des entités montrées par cette [requête SPARQL] sur Wikidata.
9. Pour scripts/ressources, voir le dossier [work] du dépôt.
10. Suivant Bernhard (2014), nous utilisons *lexème* dans le sens de *lexeme* chez Bauer (2003) : Un mot du dictionnaire ; une unité abstraite du vocabulaire, réalisée par des mot-formes représentant le lexème et sa morphologie flexionnelle. Une des formes est choisie par convention afin de nommer le lexème dans une entrée de dictionnaire ou ouvrage similaire.

2.5 Modélisation des descripteurs sociaux des personnages

Une question de modélisation qui se pose avec le corpus concerne la formalisation des variables sociales qui décrivent les personnages et les relations entre eux ; il s'agit d'attributs des personnages pertinents pour l'analyse linguistique et dramatique. Des typologies pour modéliser les personnages, formalisables en TEI, existent déjà (Galleron, 2017). Or, elles peuvent être complétées concernant la description des professions des personnages. Nous avons commencé à développer une typologie multilingue de professions avec des termes en alsacien, français et allemand (langues des professions dans les listes de personnages du corpus) ainsi qu'en anglais, car notre recherche cible un public intéressé aux langues régionales mais qui ne maîtrise pas forcément l'alsacien, et souvent anglophone. Une question de recherche est de savoir comment représenter la typologie de façon à faciliter l'interopérabilité et son intégration dans l'encodage TEI. Tant les « feature structures » (hiérarchies de caractéristiques) proposées par Galleron que le formalisme sur la base d'attributs RDFa (un format web sémantique) intégrés dans la TEI (Ruiz Fabo *et al.*, 2020) peuvent être considérés.

3 Variation et alternance codique

Le corpus présente une énorme variabilité orthographique. À ceci s'ajoute l'alternance codique entre alsacien, allemand et français. Cette section montre des exemples de chaque phénomène, qui donnent une indication des défis que pose le corpus pour des tâches de TAL comme l'identification de variantes orthographiques.

3.1 Variation

Les parlers dialectaux d'Alsace sont caractérisés par une grande variation à l'oral, qui se traduit par autant de variation à l'écrit. Dans les pièces de théâtre, la variation dans la scripturalisation dépend de la variété dans laquelle s'exprime le dramaturge, mais aussi des variations 'internes' à la pièce, c'est-à-dire relatives aux personnages, en fonction de leur origine géographique et partant, sociale. On peut citer comme cas d'école le discours prêté aux personnages présents dans le *Pfingstmontag* (1816) de J.G. Arnold, première pièce de théâtre en alsacien, dont le but est de « dresser un petit monument linguistique alsacien [11] ». Dans cette comédie, ce sont les dialectes et autres variétés linguistiques présentes en Alsace à l'époque (allemand 'standard' et français) qui sont véritablement mis en scène par leurs personnages. Dans cette pièce on retrouve des représentants de la bourgeoisie strasbourgeoise, s'exprimant dans la variété dialectale de la ville, mais aussi des représentants stéréotypés de la paysannerie du Kochersberg (une région rurale proche de Strasbourg).

La variation graphique peut donc varier d'un personnage à l'autre, comme c'est le cas chez Claus, le paysan du Kochersberg s'exprimant dans sa variante dialectale et chez Wolfgang, magister ès philosophie, s'exprimant quant à lui en allemand 'standard'. Les deux personnages emploient le verbe *(an)fragen* (questionner), ce qui donne les variations de scripturalisation dans (1) et (2). [12]

11. « [D]ie Bestimmung eines kleinen alsatischen Sprachdenkmals », comme l'exprime Arnold dans sa préface au *Pfingstmontag*.

12. La graphie qui représente la racine du verbe, sans préfixes ou suffixes, est identifiée en caractères gras. Les traductions vers le français sont données avec les exemples. Des versions encodées en TEI pour les pièces citées sont disponibles sur notre dépôt public, sauf dans le cas du *Herr Maire* (disponible sans encodage sur Numistral).

(1) I **fröau** ob err no' brüche d' Pfärd
Je demande si vous avez encore besoin des chevaux

(2) Wir sollten doch zuerst bei ihr zu Haus an**frag**en
Nous devrions d'abord aller poser la question chez elle

Dans le *Chrischtowe* de Clemens (3), ainsi que dans *Sainte Cécile* de Julius Greber (4), la racine du même verbe présente tant la graphie *frö* que *fröu* ; cette dernière est aussi trouvée dans *In's Ropfer's Apothek* par Gustave Stoskopf (5).

(3) Äi sie ruefe mr alli e Üwername. Dr Schuelmäischter hett mi g**frö**jt wie „der Ofen" häisst — no hawi gsäit „Furneau"
Ils me donnent tous un surnom. Le maître d'école m'a demandé comment on dit « der Ofen » [le four] — j'ai dit « Furneau »

(4) Do kannsch lang **fröu**je — — er saat nix, ken Wort schnüüft er
Tu peux redemander sans cesse — — Il dit rien, il ne pipe mot

(5) Ich hab e schoene Schrecke bekumme, wie 'r mich waje d'r Susanne g'**fröu**it hett
J'ai vraiment eu peur quand il m'a demandé par rapport à Susanne

Dans *D'r Herr Maire* (1898) de Stoskopf, différentes variétés sont également mises en scène et un même lexème peut à nouveau prendre des graphies divergentes à l'extrême. Dans (6), *Daö* représente l'adaptation phonographique au dialecte du Kochersberg de *Tag* (*jour*).

(6) Un dass dich guet schicksch un Savuar-Wiewr an de **Daö** leisch !
Tu as intérêt à bien te comporter et à faire preuve de savoir-vivre

Les occurrences ***Daa*** et ***Tag*** apparaissent dans la même pièce ; la première est prononcée par le fils du riche épicier strasbourgeois Pfeffer, qui s'exprime dans sa variante strasbourgeoise et la seconde apparaît dans une lettre, écrite en allemand standard, faisant également état de la diglossie médiale alors en vigueur.

3.2 Alternance codique

Le corpus présente de l'alternance codique entre variétés alsaciennes, français et allemand ; dans certains cas d'autres variétés régionales sont également présentées, comme c'est le cas de l'allemand de Saxe chez *D'r Hoflieferant* par Stoskopf, à travers le personnage Hans Grinsinger.

À part le mélange d'autres langues avec le français, une caractéristique additionnelle dans certaines pièces est l'écriture du français 'à l'alsacienne'. Dans le *Pfingstmontag*, le personnage du licencié, Alsacien âgé essayant de montrer l'étendue de ses connaissances en français, est particulier, dans la mesure où son discours est truffé de termes français, dont la prononciation est largement adaptée au dialecte alsacien, comme le révèlent les graphies dans (7a-e) :

(7a) Nong
Non

(7b) Pardong
Pardon

(7c) Wui wui
Oui, oui

(7d) Sannebawrä
Ça n'est pas vrai

(7e) Ong nangtang riäng ... Pong, Pong ... Mongtong dong sangfassong
On n'entend rien ... Bon, bon ... Montons donc sans façons

Le *Herr Maire* de Stoskopf (1898) reprend l'idée de transcrire le 'français-alsacien' déjà utilisé par Arnold en 1816 : On le voyait dans l'expression *Savuar-Wiewr* pour *savoir-vivre* de l'exemple (6) ci-dessus.

D'r Hoflieferant de Stoskopf (1905) est un autre exemple des subtilités qui peuvent être représentées dans le corpus concernant l'alternance de variantes. Dans cette pièce, les personnages utilisent parfois la prononciation française ou allemande des noms de famille pour exprimer leur identité et leur proximité à leur interlocuteur ou leur rejet de celui-ci ; l'utilisation de la prononciation française est alors indiquée en italiques, comme dans l'exemple suivant par le personnage Fritz Grinsinger :

(8) *Pardon*, dass ich Sie unterbrech, erschtens bin ich noch lang nit Ihr Liewer und zweitens heiss ich nit Grinsinger [avec prononciation allemande], ich heiss *Grinsinger* [avec prononciation française, en italiques dans l'original].
Pardonnez mon interruption, mais premièrement je ne suis pas votre cher [monsieur Grinsinger] et deuxièmement je ne m'appelle pas Grinsinger, mais *Grinsinger*.

Comme le montrent les exemples dans cette section, le corpus va au-delà de cas 'simples' d'alternance codique. Nous prévoyons une représentation TEI basique du phénomène avec des éléments `<seg>` et des attributs `@xml:lang`, comme vu en (2.4). La possibilité d'encoder plus de détails (ce qui serait évidemment permis par le standard TEI) est une question ouverte. La détection automatique des cas d'alternance codique est un autre sujet de recherche possible sur le corpus.

4 Identification automatique de variantes orthographiques

La neutralisation des variantes est incontournable pour comparer le contenu des pièces et faire des analyses thématiques, p. ex avec le *topic modeling* (Blei, 2012) ou des méthodes de textométrie (Lebart *et al.*, 2019) ; ces méthodes demandent une représentation orthographique homogène du vocabulaire. En outre, cette neutralisation pourra bénéficier à la recherche en texte intégral, une fonction de base de l'interface d'exploration du corpus qui sera développée dans la phase finale du projet.

Ces questionnements ne sont pas nouveaux et deux approches différentes peuvent ici être envisagées : soit les variantes sont normalisées vers une forme correspondant à une norme choisie, soit elles sont tout simplement reconnues comme étant des variantes, sans qu'il y ait pour autant une normalisation explicite.

La normalisation orthographique automatique, en tant que tâche de Traitement Automatique des Langues (TAL) a notamment été appliquée pour l'analyse de textes du web social (Han et Baldwin, 2011; Alegria *et al.*, 2015; Doval *et al.*, 2020). Dans ce cas précis, la normalisation des mots hors vocabulaire (fautes d'orthographe, orthographe non conventionnelle, abréviations) se fait généralement vers la forme standard. La normalisation est aussi utilisée pour les variétés historiques (Etxeberria *et al.*, 2016; Bollmann *et al.*, 2017; Bollmann, 2019). La norme est alors souvent la forme standard contemporaine, même si cela pose la question des formes disparues, qui n'ont pas d'équivalent dans la variété contemporaine.

La deuxième approche consiste à identifier les variantes sans chercher pour autant à les normaliser : en effet, pour de nombreuses applications, comme la recherche dans un corpus, la normalisation n'est pas nécessaire. Il s'agira ainsi de repérer les variantes, par exemple à l'aide de méthodes non

supervisées de *clustering* (Dasigi et Diab, 2011; Rafae *et al.*, 2015) ou des méthodes supervisées qui déterminent si deux formes sont des variantes ou non (Barteld *et al.*, 2019). Nous nous orientons également vers ce type d'approche, pour faire suite à de premières expériences visant à identifier les variantes dans des lexiques bilingues alsacien-français (Bernhard, 2014) [13]. Il n'y a en effet pas de "norme" orthographique stable à laquelle nous pourrions nous référer pour les dialectes alsaciens. Même si l'allemand est souvent considéré comme la forme écrite à privilégier pour l'alsacien, cela ne reflète pas la réalité de nos corpus, comme nous avons pu le montrer dans la section précédente.

Nous testons actuellement des méthodes de classification supervisée (cf. Barteld *et al.*, 2019) et les résultats sont en cours d'analyse. Nous aimerions échanger avec la communauté sur des approches permettant de profiter au mieux d'un nombre limité de données d'entraînement, avant de nous engager dans la création de nouvelles données (annotées) pour la tâche.

5 Perspectives

Après avoir encodé les premières pièces du corpus, plusieurs intérêts de recherche, en partie évoqués *supra*, sont les suivants : d'un côté, implémenter la modélisation TEI des variables sociales décrivant les personnages. D'un autre côté, l'application possible de méthodes d'apprentissage automatique à la détection des éléments structurels des pièces (répliques, didascalies) pour leur encodage TEI automatique. Finalement, nous sommes en train d'évaluer l'application de méthodes de TAL à l'identification automatique de variantes, ce qui constituerait un bon apport à l'exploitabilité du corpus pour des analyses linguistiques et de contenu. En outre, la FAIRisation du corpus sera complétée par sa mise à disposition sur des plateformes ouvertes d'exposition de données.

Remerciements

Ce travail a bénéficié d'un financement dans le cadre de l'IdEx Université de Strasbourg. Nous remercions également les stagiaires ayant participé à l'encodage des pièces : Audrey Deck et Soihira El-Kabir. Merci aux relecteur·trice·s pour leurs commentaires détaillés qui ont aidé à améliorer l'article.

Références

ALEGRIA, I., ARANBERRI, N., COMAS, P. R., FRESNO, V., GAMALLO, P., PADRÓ, L., SAN VICENTE, I., TURMO, J. et ZUBIAGA, A. (2015). TweetNorm : a benchmark for lexical normalization of Spanish tweets. *Language Resources and Evaluation*.

BARTELD, F., BIEMANN, C. et ZINSMEISTER, H. (2019). Token-based spelling variant detection in Middle Low German texts. *Language Resources and Evaluation*, pages 1–30.

BAUER, L. (2003). *Introducing linguistic morphology*. Edinburgh University Press Edinburgh.

13. Les habitudes de scripturalisation du corpus (utilisation du eszett par exemple, ou utilisation du graphème simple <u> pour rendre le <ou> français) sont obsolètes par rapport aux pratiques actuelles, ce qui demande l'adaptation des méthodes ; des ressources pour le TAL en alsacien ont été développées par le projet ANR RESTAURE (Bernhard *et al.*, 2019) mais un corpus diachronique de théâtre demande d'élargir les ressources.

BERNHARD, D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources : the Example of Alsatian. *In Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, pages 23–29, Reykjavík, Iceland.

BERNHARD, D., BRAS, M., ERHART, P., LIGOZAT, A.-L. et VERGEZ-COURET, M. (2019). Language Technologies for Regional Languages of France : The RESTAURE Project. *In International Conference Language Technologies for All (LT4All) : Enabling Linguistic Diversity and Multilingualism Worldwide*, Paris, France.

BLEI, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

BOLLMANN, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

BOLLMANN, M., BINGEL, J. et SØGAARD, A. (2017). Learning attention for historical text normalization by learning to pronounce. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 332–344, Vancouver, Canada. Association for Computational Linguistics.

DASIGI, P. et DIAB, M. (2011). CODACT : Towards Identifying Orthographic Variants in Dialectal Arabic. *In Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 318–326, Chiang Mai, Thailand.

DOVAL, Y., VILARES, J. et GÓMEZ-RODRÍGUEZ, C. (2020). Towards robust word embeddings for noisy texts. *arXiv :1911.10876 [cs]*. arXiv : 1911.10876.

ETXEBERRIA, I., ALEGRIA, I., URIA, L. et HULDEN, M. (2016). Evaluating the Noisy Channel Model for the Normalization of Historical Texts : Basque, Spanish and Slovene. *In LREC*.

FISCHER, F. et BÖRNER, I. (2019). Programmable Corpora : Introducing DraCor, an Infrastructure for the Research on European Drama. *In Digital Humanities 2019*, page 5, Utrecht.

GALLERON, I. (2017). Conceptualisation of Theatrical Characters in the Digital Paradigm : Needs, Problems and Foreseen Solutions. *Human and Social Studies*, 6(1):88–108.

HAN, B. et BALDWIN, T. (2011). Lexical Normalisation of Short Text Messages : Makn Sens a #twitter. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

HUCK, D. (2015). *Une histoire des langues de l'Alsace*. La Nuée Bleue.

HUCK, D., BOTHOREL-WITZ, A. et GEIGER-JALLET, A. (2007). L'Alsace et ses langues. Eléments de description d'une situation sociolinguistique en zone frontalière. *In Aspects of Multilingualism in European Border Regions : Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, pages 13–101. EURAC Research (Europäische Akademie / Accademia Europea / European Academy), Bozen/Bolzano.

KHEMAKHEM, M., FOPPIANO, L. et ROMARY, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *In electronic lexicography, eLex 2017*, Leiden, Netherlands.

KHEMAKHEM, M., ROMARY, L., GABAY, S., BOHBOT, H., FRONTINI, F. et LUXARDO, G. (2018). Automatically Encoding Encyclopedic-like Resources in TEI.

LEBART, L., PINCEMIN, B. et POUDAT, C. (2019). *Analyse des données textuelles*. Presses de l'Université du Québec, 1 édition.

RAFAE, A., QAYYUM, A., MOEENUDDIN, M., KARIM, A., SAJJAD, H. et KAMIRAN, F. (2015). An Unsupervised Method for Discovering Lexical Variations in Roman Urdu Informal Text. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 823–828.

RUIZ FABO, P., BERMÚDEZ SABEL, H., MARTÍNEZ CANTÓN, CLARA et GONZÁLEZ-BLANCO, ELENA (2020). The Diachronic Spanish Sonnet Corpus (DISCO) : TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings. *Digital Scholarship in the Humanities*.

TEI CONSORTIUM (2020). TEI P5 : Guidelines for Electronic Text Encoding and Interchange. Publisher : Zenodo.

WAGNER, A. (2020). TEI XML to Zenodo service published : Automatic depositing the project's TEI files at a long-term archive – Die Schule von Salamanca.

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., AXTON, M. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3(1).

D'un corpus à l'autre

D'une étude reproductible et portable du discours direct nisvai à la comparaison linguistique

Jocelyn Aznar
(1) ZAS, Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
Schützenstr. 18,10117 Berlin, Allemagne
(2) Aix-Marseille Université, CNRS, EHESS, CREDO UMR 7308, 3, Place Victor Hugo,
13331 Marseille, France
aznar@leibniz-zas.de

RÉSUMÉ

À partir d'une étude comparative en cours portant sur le discours direct à travers des documentations linguistiques de langues orales peu documentées, nous proposons une réflexion sur la reproductibilité et la portabilité d'une recheche en linguistique. L'enjeu est de porter l'étude du discours direct réalisées sur le corpus de narrations nisvaies, une langue orale du Vanuatu, à d'autres corpus de langues orales. Les annotations de ces corpus ont été amendés et normalisées par les efforts combinés des projets DoReCo et QUEST. Nous verrons que si la reproductibilité d'une étude sur une langue facilite sa critique, la question de la portabilité d'une étude vers d'autres corpus requiert que ces derniers répondent à des normes et unités interopérables aussi bien d'un point de vue informatique que linguistique.

ABSTRACT

From one corpus to another: From a reproducible and portable study of nisvai direct discourse to linguistic comparison

Based on an ongoing comparative study of direct discourse through poorly documented linguistic documentation of oral languages, we propose a reflection on the reproducibility and portability of research in linguistics. The challenge is to bring the study of direct discourse carried out on the corpus of Nisvai narratives, an oral language of Vanuatu, to other corpuses of oral languages. The annotations of these other corpora have been corrected and standardised by the combined efforts ofthe DoReCo and QUEST projects. We will see that while the reproducibility of a study on a language facilitates its criticism, the question of the portability of a study to other corpora requires that the latter meet standards and units that are interoperable from both a computational and linguistic point of view.

MOTS-CLÉS : comparaison linguistique, langues orales, documentation linguistique, corpus de terrain, portabilité, reproductibilité.

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 42–52,

KEYWORDS: cross-linguistics, oral languages, linguistic documentation, eldwork corpora, annotations, portability, reproductibility.

1 Introduction

La présente communication est une réflexion sur l'étude linguistique de corpus de langues orales à travers des techniques associées aux sciences des données. Le propos est de montrer comment les concepts de reproductibilité et de portabilité sont appliqués à l'étude de corpus de langues orales. Il s'agit de mettre en place des pratiques et des méthodes qui facilitent la critique, le partage et la réutilisation des données et des résultats d'une étude.

Cette réflexion, ainsi que l'étude sur laquelle elle s'appuie, sont issues de QUEST, un projet de recherche financé par le ministère des Sciences allemand afin de promouvoir et valider des normes et bonnes pratiques pour élaborer de corpus de documentation linguistique ré-utilisables. Plus particulièrement, cette étude s'insère dans le sous-projet RefCo, une initiative dont l'objectif est de mettre à disposition des corpus de référence adaptée à la comparaison linguistique. Les corpus de documentation linguistique utilisées pour cette étude sont ceux qui ont été amendés par les projets DoReCo (Paschen et al. 2020) et QUEST (https://cutt.ly/quest-project). L'étude porte sur la réalisation du discours direct au sein de ces corpus, en commençant par le corpus de narrations nisvaies (Aznar 2019) afin mettre au point une base de références techniques, pour s'étendre à un ensemble de cinq corpus documentant les langues beja, mojeno, sanzhi dargha et arapaho, corpus qui ont été également corrigés par DoReCo et QUEST. C'est dans ce cadre que s'inscrit la réflexion sur les concepts de reproductibilité et de portabilité : la comparaison linguistique de langues orales à travers des documentations réalisées avec le logiciel ELAN, dont les textes sont des narrations monologiques et dont les annotations ont été vérifiées et corrigées par DoReCo et QUEST.

Un premier point à aborder avant de rentrer dans le cœur du propos est la terminologie. Les termes portabilité et reproductibilité sont sujets à débat dans de nombreuses disciplines scientifiques[1], l'enjeu n'est ici pas de proposer des définitions conceptuelles précises mais de réfléchir à leur implémentation pratique dans un cadre précis. L'acceptation de portabilité s'inspire du sens proposé par Bird et Simons (2003) que l'on peut résumer comme la mise en place d'un ensemble de moyens et de pratiques permettrant de réutiliser une documentation linguistique sur le long terme. Le concept est appliqué ici à l'étude des données linguistique et non à la documentation. Il s'agit de réfléchir aux pré-requis qui sont à considérer afin qu'une étude linguistique assistéee par des traitements automatisés sur un corpus puisse être réalisée sur d'autres corpus. La portabilité de l'étude du discours direct s'inscrit donc dans une démarche de reproductibilité et de réplicabilité de la recherche. Reproductibilité est entendu ici comme la possibilité de répéter une étude est fourni

[1] Les définitions des concepts de réplicabilité, reproductibilité, répétabilité varient en fonction de la langue où le concept est défini, de la discipline ou de la position théorique de l'auteur. Il ne s'agit pas ici de faire une proposition définitive quant à ses termes mais simplement de mieux expliciter notre proposition (voir par exemple (Berez-Kroeker et al. 2018; Drummond 2009; Rey-Coyrehourcq et al., s. d.; McArthur 2019).

aux lecteurs à travers l'accès aux données et aux algorithmes nécessaires pour arriver au résultat décrit. Quant à la replicabilité, elle est définie comme la possibilité de refaire une même étude dans un contexte différent afin d'obtenir des résultats permettant de discuter l'étude de référence.

Maintenant que nous avons vu le contexte dans lequel s'insère cette comparaison des discours direct, nous allons aborder différentes étapes de l'étude algorithmique à travers le logiciel Jupyter pour ensuite décrire différents problèmes auxquels nous sommes confrontés lorsque nous souhaitons porter cette étude d'un corpus à l'autre.

2 Étudier le discours direct au sein du corpus Nisvai reproductible avec Jupyter

La présente description repose sur le corpus de pratiques narratives nisvaies, un corpus d'enregistrements audio annoté avec le logiciel ELAN (Aznar, 2019). Dans le cadre de mon travail au sein de QUEST, le corpus a été soumis à l'évaluation de DoReCo et RefCo. À cette fin, les enregistrements audio et leurs annotations ont été rendus accessibles sous licence Creative Commons BY-NC-ND, une licence qui permet de partager et d'étudier ces données.

2.1 QUESTCorpora: un module python pour créer des données linguistiques à partir des annotations du corpus

Afin de fournir aux corpus de QUEST une interface informatisée unifiée et partageable, je développe un module python nommé QUESTCorpora qui permet d'accéder via des objets pythons à ses différents composants : Corpus, Textes, Annotations. L'enjeu du module est de produire des données linguistiques comparables à partir des corpus.

Pour ce faire, le module s'appuie sur un ensemble de corpus gérés par le projet RefCo qui ont été créés et retravaillés de manière suffisamment similaires pour en extraire des données linguistiques comparables. La production de ces données linguistiques comparables passent toutefois par une politique de gestion des annotations. Cette politque se traduit par un ensemble de paramètres dédiés à chacun des corpus qui configurent les algorithmes de gestion des annotations.

Concrètement, QUESTCorpora inspecte les fichiers ELAN des corpus gérés par RefCo, les transformes en instances Python pour fournir une interface permettant de manipuler les annotations et de produire des informations à travers des traitements automatisés (expresssions régulières, calculs et statistiques) sur chacune des annotations. Les résultats provenant de ces traitements sont ensuite transformés en un tableau de données afin d'être manipulables via les modules Python dédiés aux traitements des données (voir notamment Pandas, Statsmodels). Ces modules permettent de produire des représentations statistiques et visuelles de ces annotations et des résultats des traitements automatiques.

2.2 Jupyter pour faciliter la reproductibilité d'une étude de corpus

Dans les domaines impliquant une chaine de traitements informatisés des données, l'utilisation d'environnement de partage documenté des algorithmes employés se développent (Stodden, Leisch, et Peng 2014). L'étude comparée des discours directs repose en partie sur des traitements automatisés des corpus de données linguistiques. L'utilisation de Jupyter (Thomas et al. 2016), un « carnet » en ligne, permet d'annoter les différentes étapes d'une chaine de traitements informatiques dans un format qui facilite leurs partage sous la forme d'une narration. D'un côté, le logiciel aide le chercheur lorsqu'il explore, teste ses hypothèses sur ses données ou travaille en équipe, du côté des relecteurs, il facilite la critique et l'appropriation des étapes ayant conduit aux résultats. L'utilisation d'un module Python pour intérargir avec les corpus facilite l'intégration de ces données dans une chaine de traitements automatisés, reproductible et partageable.

Import des modules nécessaires à cette étude:

```
import pandas as pd # help handling the data
import matplotlib.pyplot as plt # for representing the data
import numpy as np # not used fro now, just in case
import statsmodels.formula.api as sm # not used yet, just in case
import seaborn as sns # to produce charts and visualizations,
import re # to parse the texts and segments of textual annotation
import os
```

```
import QuestCorpora
```

Analyse du corpus

```
#Creation du corpus de données linguistiques

Study = QuestCorpora.Study("pour_LIFT", "22.11.2020", "list_texts.csv", "corpus_LIFT_RefCo.csv")
Study.get_generic_data()
Study.re_analysis("(dire|parler)", "morphology", '_dire')
Study.re_analysis(":", "transcription", "_deuxpoints_transcription")
Study.re_analysis(":$", "transcription", "_deuxpointsFin_transcription")
Study.re_analysis(":.+«[^»]+$", "transcription", "_deuxpointsFin_transcription")
Study.re_analysis("[^«]+»", "transcription", "_uniquFin_transcription")
Study.re_analysis("«.+»", "transcription", "_lesDeux_transcription")
Study.re_analysis("[^»]+«.+»[^«]+", "transcription", "_monoUniquement_transcription")
Study.re_analysis(":.+:", "transcription", "_Multi_transcription")
Study.re_analysis(":", "translation", "_deuxpoints_traduction")
Study.write_data("Quest_Nisvai_DirectSpeech.csv", "csv")
```

```
# Import des données dans Pandas
datafile = "Quest_Nisvai_DirectSpeech.csv"
data = pd.read_csv(datafile)
```

Figure 1: Exemples d'étapes pour la production de données linguistiques à partir du module QUESTCorpora

La figure 1 montre un exemple de chaine de traitements automatisé réalisé sur un corpus du projet QUEST, le corpus nisvai. La première étape consiste en l'import des modules qui seront sollicités dans le cadre de l'étude. La deuxième étape consiste en l'instantiation d'une étude sur corpus à travers des expressions régulières. Dans le cadre de cette étude, des recherches à l'aide d'expressions régulières sont appliquées à chacune des unités d'annotation du corpus. Si la recherche est concluante, elle retourne un 1 dans la ligne du tableau[2] correspondant à l'annotation, dans le cas contraire, un zéro est retourné.

La figure 2 présente quelques colonnes contenant les données du corpus nisvai intégré dans le module QUESTcorpora. Les deux dernières colonnes correspondent aux résultats des requêtes réalisées avec des expressions régulières sur les annotations. Ce sont à partir de ces colonnes contenant soient les données des annotations du corpus, soit les métadonnée associées aux

[2] Il s'agit ici plus précisément d'un objet DataFrame proposé par le module Python Pandas (https://pandas.pydata.org/).

	index	time_1	time_2	reference	transcription	morphology	translation	coherency	_dire	_deuxpoints_transcription
0	nisvai_25_174	589390	590480	T41.2015.174	Kusvai. -- Kusvai.	--	Kousvé. -- Kousvé.	False	0	0
1	nisvai_25_173	587480	589060	T41.2015.173	Nabol nyn ga=qan nyn ni, ga=cub urun.	histoire DEI.R 3SG=être_comme DEI.R INT 3SG=te...	Cette histoire est comme cela, elle se termine...	True	0	0
2	nisvai_25_172	581170	586240	T41.2015.172	Ga=hub nabu-n naho-n wantaim, nahemac ili ga=m...	3SG=lapider ??-3SG face-3SG INT démon DET 3SG=...	Il la jette en pleine face d'un coup, le mécha...	True	0	0
3	nisvai_25_171	574770	580350	T41.2015.171	Naremac ili ga=kal : «Ui, asi na=han-i?» Ga=ka...	démon DET 3SG=dire INTER qui 1SG=manger-INTR 3...	Le méchant dit : «Eh, qui est-ce que j'ai mang...	False	1	1
4	nisvai_25_170	572560	573550	T41.2015.170	Haiq qa=han a=nanaq sa-q!»	2SG 2SG=manger ART.P=mère PA-1SG	Tu as mangé ma mère !»	True	0	0

Figure 2: Extrait d'un tableau Pandas de données d'un corpus QUESTCorpora dans l'interface de Jupyter.

locuteurs, telles que l'âge, le sexe, le genre du texte, ou au texte et à sa situation de production, genre discursif, date, corpus ou enfin les résultats des requêtes, que sont produites les analyses sur les corpus.

3 D'un corpus à la comparaison de corpus

Maintenant que nous avons vu les étapes d'une étude informatisée reproductible sur un corpus, nous devons nous intéresser au portage d'une étude vers un autre corpus.

La comparaison s'appuie sur l'utilisation de corpus ouverts, c'est-à-dire dont la licence d'utilisation permet au moins la copie et la consultation. Toutefois, si l'ouverture des corpus participe à la portabilité d'une étude, cette qualité n'est pas suffisante pour permettre la portabilité. Il est également nécessaire que les annotations de ces corpus soient documentées et que les unités linguistiques annotées soient comparables. Enfin, que ce soit pour la reproductibilité ou la portabilité de l'étude, il est nécessaire d'avoir accès à la méthode qui a permis d'arriver aux résultats. Lorsqu'un article linguistique décrit les résultats issus d'une étude et leur interprétation, les algorithmes employés pour réaliser l'étude ne sont pas nécessairement fournis, en particulier lorsque les algorithmes ne sont pas informatisés. C'est la combinaison de ces qualités, d'un côté des données accessibles et normalisés, et de l'autre des algorithmes ouverts, qui permet la mise en place d'études reproductibles et portables.

3.1 Les étapes du portage d'une étude à d'autres corpus

Afin de pouvoir adapter le comportement du module QUESTCorpora aux spécificités de chacun des corpus, chaque corpus est associé à un fichier de configuration spécifique[3]. Chaque corpus doit être étudié en fonction de l'étude. Dans notre cas, il s'agit d'étudier le discours direct, et plus particulier les différents silences qui entourent sa mise en scène. Dans cette section sont rapportées les différentes sources de difficultés quant à l'observation du discours direct de manière automatisée à travers les corpus.

3.2 L'indexation temporelle par rapport à l'enregistrement audio

[3] Les chiers de con guration sont écrits en TOML (voir https://toml.io/en/), un format de chier qui vise à être lisible tout en étant analysable d'un point de vue informatique.

La première source de difficulités sont les variations d'un corpus à l'autre au niveau de l'indexation des unités d'annotation par rapport aux enregistrements audio. Il s'agit de la stratégie qui a été mise en place lors de la transcription afin de segmenter le flot de la parole en unités d'annotation. Une première variation provient des possibilités offertes par ELAN, qui permet de produire des annotations soit continues, soit discontinues, c'est-à-dire soit collées les unes aux autres ou au contraire séparées par temps non annotés[4]. Plusieurs pratiques ont été relevées au sein des corpus DoReCo-QUEST :

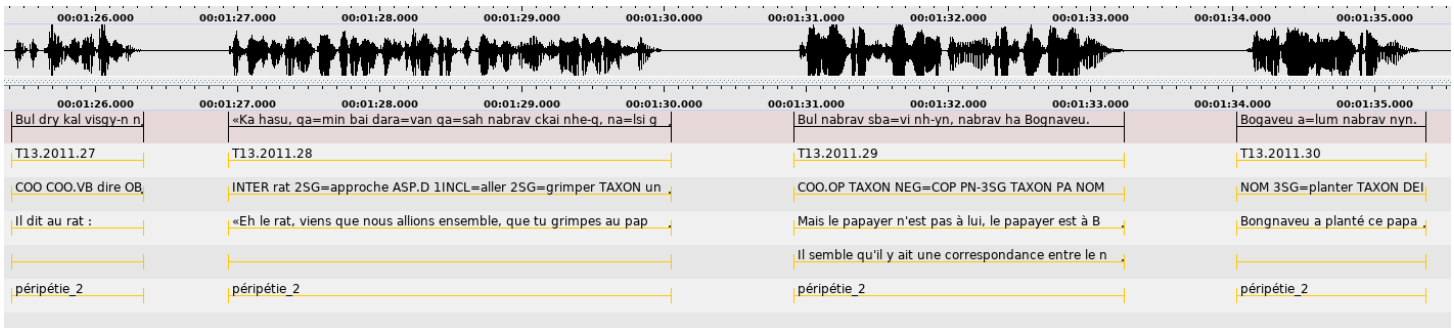

Figure 3: Exemple d'indexation disccontinue des unités d'annotation, avec une unité sur deux ne contenant qu'un silence.

- Une indexation par rapport aux groupes de souffles, voir figure 3 : les unités d'intonation correspondent aux moments où le locuteur parle, lorsqu'il marque une pause, que ce soit pour reprendre son souffle, structurer son texte ou simplement dans le cadre d'une dysfluence (réflexion ou hésitation), alors une segmentation est opérée.
- Une indexation syntaxico-sémantique, voir figure 5: le linguiste segmente en fonction d'une unité linguistique qu'il identifie, potentiellement avec l'aide d'un locuteur de la langue. Cette segmentation correspond le plus souvent à une segmentation en phrases. Elle peut prendre en considération les pauses marquées par les locuteurs pour borner les unités d'annotation, mais ses unités d'annotation peuvent comprendre des pauses en leur sein.
- Une indexation continue, voir figure 4, produisant des unités d'annotation aussi bien pour les temps de paroles que pour les temps de silences.
- Une annotation continue, voir figure 6, intégrant les temps de silence dans unité d'annotation au sein de ses bornes.

[4] Nous verrons par la suite que la non-annotation peut être considéré comme une forme de silence.

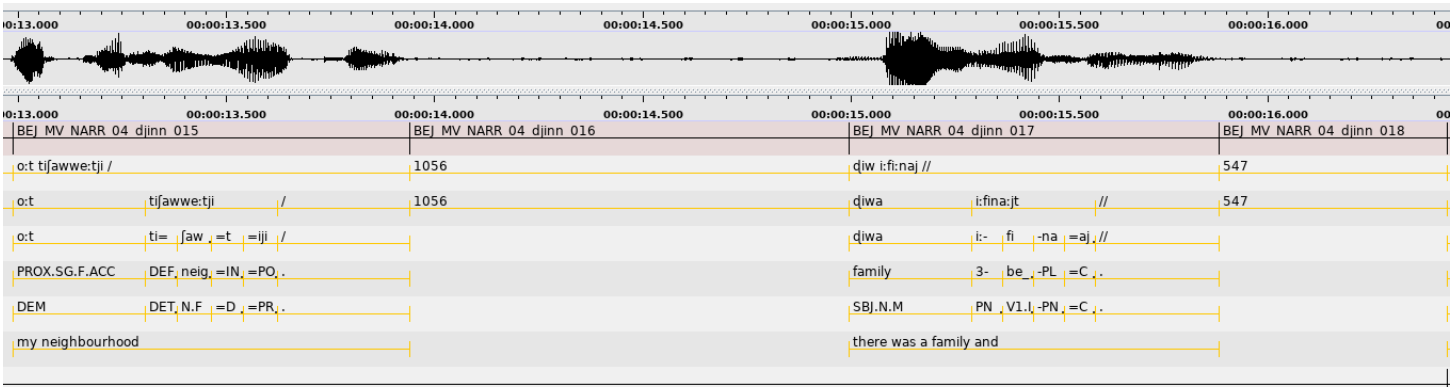

Figure 4: Exemple d'indexation continue des unités d'annotation, avec une unité sur deux ne contenant qu'un silence.

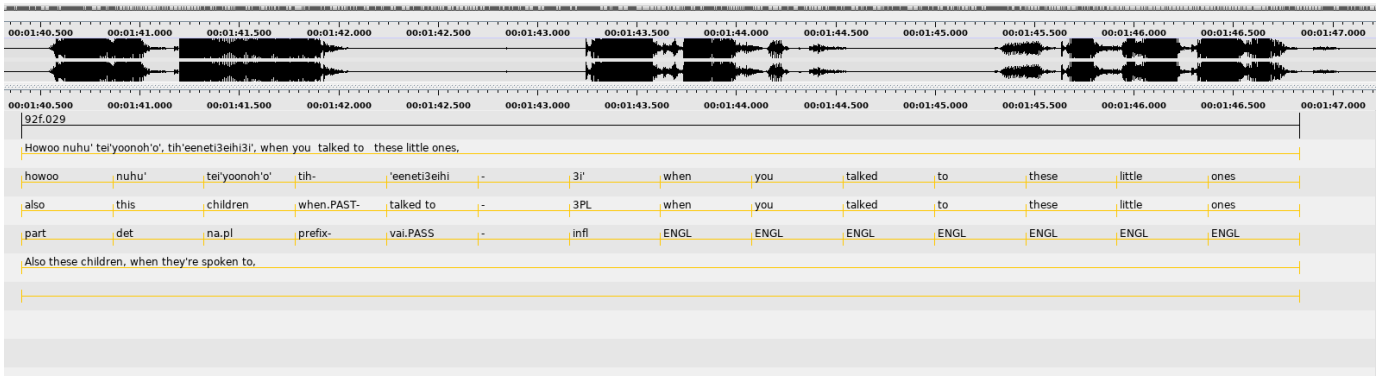

Figure 5: Exemple d'indexation discontinue avec des silences absorbés au sein de l'annotation et dont les limites ne sont que peu prises en compte.

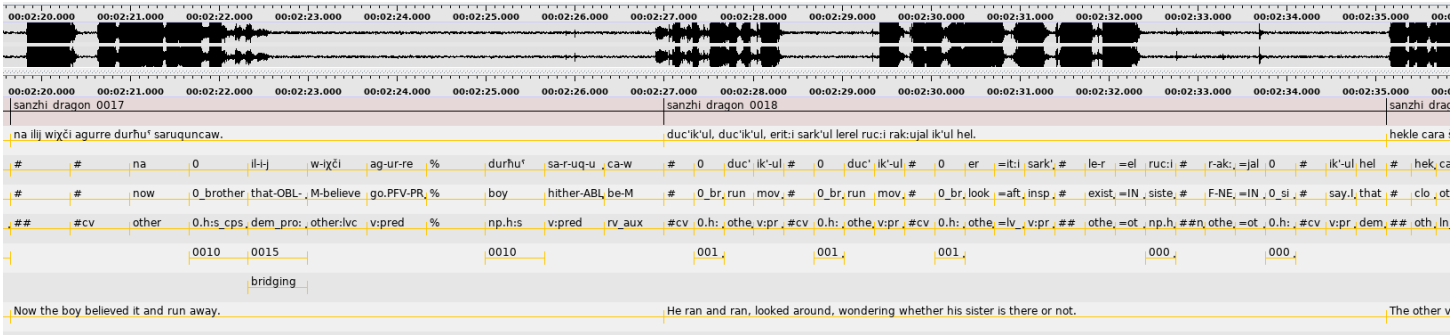

Figure 6: Exemple d'indexation continue des unités d'annotation avec intégration du temps de silence dans la suite de l'unité d'annotation

Dans le cadre de l'étude du silence de la réalisation du discours direct, deux paramètres sont identifiables afin de distinguer les pratiques des linguistes : la continuité ou non des unités d'annotation et la prise en compte du silence dans les unités d'annotation

	Discontinuité	Continuité
Prise en compte du silence	Silence représenté par la non-annotations	Silence correspondant à des annotations
Non prise en compte	Silence rogné par les unités d'annotation	Silence intégré avec le groupe de parole le précédent

À noter qu'une autre pratique a été réalisé par MAUS (Strunk, Schiel, et Seifart 2014), le système employé par DoReCo-QUEST afin d'indexer les graphèmes de la transcription à l'enregistrement. Dans ce cas, le logiciel indique par la balise <p:> les silences au sein d'un tier dont la segmentation est continue.

3.3 La transcription et l'interprétation des systèmes d'écriture

Parmi les corpus gérés par les projets DoReCo et QUEST, si à un texte est associé une couche de transcription, il est également possible dans certains corpus que plusieurs couches se rapportent à la transcription. Dans ce cas, il s'agit . L

Les systèmes d'écriture utilisés pour transcrire les différentes langues n'ont pas les mêmes valeurs. Ceci est résolu à l'aide d'un tableau associant les graphèmes employés par les linguistes à leurs valeurs phonétiques ou phonologiques.

En ce qui concerne la représentation du discours direct au sein de la transcription, un certain nombre de choix ont été opérés par les transcripteurs. En fonction de la langue maternelle de la personne ayant réalisée la transcription, les conventions typographiques pour représenter le discours direct peuvent variée : utilisation, ou non, des guillemets français ou anglais ; marquage du début du discours direct par deux points, transcription ou non de l'interjection employée par le locuteur. Ces variations ont pour conséquences que le discours direct ne puissent pas être identifié de manière systématique à travers la transcription.

3.4 Les traductions

La traduction de textes oraux est un aspect central de la documentation linguistique, mais elle ne fait l'objet finalement que de très peu de discussions de la part des linguistes travaillant sur des langues orales. La recommandation la plus courante est de traduire dans une des langues véhiculaires internationales, ou tout au moins, une langue plus connue que la langue documentée. Ce manque de discours sur la pratique explique que les textes ne soient pas traduits selon les mêmes principes en fonction des linguistes ou d'un projet à l'autre. Les destinataires des traductions ne sont pas les mêmes d'un projet à l'autre : produire une ressource bilingue pour l'école locale, annoter pour des linguistes. C'est en partie à travers l'identification de ces destinataires qu'un nombre de choix de traduction peuvent être réalisés : explicitation ou non du vocabulaire, la traduction systématique des termes, utilisation d'une typographie spécifique à un genre littéraire de destination, etc.

Du point de vue de DoReCo et de QUEST, la traduction est la couche la plus difficile à contrôler. S'il est demandé à la personne fournissant le corpus d'indiquer la langue de traduction et de fournir

un glossaire des termes de la langue source, vérifier l'adéquation entre ce qui est dit dans la langue source et ce qui est retranscrit dans la langue cible relève des compétences du linguiste sur le terrain et des personnes avec lesquelles il ou elle travaille.

En ce qui concerna l'étude du discours direct à travers la traduction, il apparait au sein des corpus DoReCo-QUEST que si la couche de transcription ne représente pas systématiquement le discours direct, celui-ci est presque systématiquement représenté au sein de la couche associée à la traduction. La couche de traduction est un moyen plus fiable pour identifier les unités d'annotation contenant du discours direct. Certains corpus possèdent plusieurs couchent de transcription qui se différencient alors au niveau de leur portée. Ainsi, pour le Beja, une première couche de traduction propose une correspondance 1:1 à la segmentation en unités d'annotation alors qu'une deuxième couche de traduction englobe plusieurs unités d'annotation afin de permettre une traduction moins littérale. Enfin, en fonction des projets dans lesquels les corpus ont été produits, les transcriptions ont pu être traduites en plusieurs langues véhiculaires.

3.5 L'annotation morphologique

Une ou plusieurs couches d'annotation morphologique sont présentes dans chacun des corpus DoReCo-QUEST. Si les annotations morphologiques sont suffisamment formalisées d'un point de linguistique afin de pouvoir faciliement associer visuellement un morphème à sa glosse, de nombreux corpus n'ont pas été annotés de manière systématique. Cela entraine des difficultés dans les traitements automatisées des annotations. Ainsi, dans certains corpus, les linguistes peuvent indiquer des synonymes aux termes décrits en utilisant plusieurs conventions typographiques au sein d'un même texte. Il est alors nécessaire de les identifier, voire de réduire ses différentes conventions à un seul et unque procédé lorsque nous sommes certain qu'il s'agit d'un même et unique sens.

Un autre problème est l'hétérogénéité des couches d'annotation de la morphologie. Certains corpus possèdent plusieurs couches contenant des informatiques liées à la morphologie de la langue. La comparaison de ce contenu en devient alors plus difficile car les couches d'annotation ne sont pas équivalentes les unes aux autres. La solution retenue est de créer une couche d'abstraction contenant les différentes annotations afin de pouvoir comparer les informations morphologiques d'une langue à l'autre.

4 Conclusion

Le portage d'une étude assistée par des traitements automatisés sur le discours direct à travers des corpus de langues orales requiert de prendre en compte les particularités de chacun de ces corpus. Il apparait toutefois que ces particularités ne sont pas infinies et peuvent être identifiées afin de faire l'objet d'une gestion des annotations dédiées afin d'obtenir des données linguistiques comparables. C'est ce qui est mis en place à travers le module Python QUESTCorpora. Afin d'éviter de réinventer un outil qui existe déjà, la réutilisation de mutiltool, actuellement pris en charge par Delafontaine (https://github.com/DoReCo/multitool) dans le cadre du projet DoReCo, est en cours d'évaluation.

Nous conclurons sur le fait que si la reproductibilité d'une étude est un critère de validation d'une étude, lorsque cette reproductibilité est combinée avec des corpus répondant aux mêmes normes, nous avons alors la possibilité de porter les études d'un corpus à l'autre afin de procéder à des comparaisons linguistiques. Finalement, en abordant la question de la portabilité d'une étude, lorsque cette problématique est combinée à la répétabilité d'une analyse algorithmique informatisée telle qu'offerte par le logiciel Jupyter, l'objectif est alors d'arriver à une réplicabilité typologique.

Références

Aznar, Jocelyn. 2019. « Narrer une nabol : La production des textes nisvais en fonction de l'âge et de la situation d'énonciation, Malekula, Vanuatu ». Marseille: EHESS.

Berez-Kroeker, Andrea L, Lauren Gawne, Susan Smythe Kung, Barbara F Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. « Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field ». *De Gruyter*, Linguistics, 56 (1): 18. https://doi.org/10.1515/ling-2017-0032.

Bird, Steven, et Gary F. Simons. 2003. « Seven Dimensions of Portability for Language Documentation and Description ». *Language*, Language, 79 (3): 557-82. https://doi.org/10/d95m65.

Drummond, Chris. 2009. « Replicability Is Not Reproducibility: Nor Is It Good Science ». *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26 Th ICML*, 4.

McArthur, Sally L. 2019. « Repeatability, Reproducibility, and Replicability: Tackling the 3R Challenge in Biointerface Science and Engineering ». *Biointerphases* 14 (2): 3. https://doi.org/doi.org/10.1116/1.5093621.

Paschen, Ludger, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, et Frank Seifart. 2020. « Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo) ». *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* 12: 2657–2666.

Rey-Coyrehourcq, Sébastien, Robin Cura, Laure Nuninger, Julie Gravier, Lucie Nahassia, et Ryma Hachi. s. d. « Vers une recherche reproductible dans un cadre interdisciplinaire: enjeux et propositions pour le transfert du cadre conceptuel et la réplication des modèles », 25.

Stodden, Victoria, Friedrich Leisch, et Roger D. Peng, éd. 2014. *Implementing Reproducible Research*. The R Series. Boca Raton: CRC Press.

Strunk, Jan, Florian Schiel, et Frank Seifart. 2014. « Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora Using WebMAUS ». *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3940--3947.

Thomas, Kluyver, Ragan-Kelley Benjamin, Pérez Fernando, Granger Brian, Bussonnier Matthias, Frederic Jonathan, Kelley Kyle, et al. 2016. « Jupyter Notebooks – a Publishing Format for Reproducible Computational Work ows ». *Stand Alone*, 87–90. https://doi.org/10/gf48c9.

Language Identification of Guadeloupean Creole

William Soto
Université de Lorraine, Nancy, France
williamsotomartinez@gmail.com

RÉSUMÉ

L'identification automatique de la langue est une étape de pré-traitement particulièrement utile lorsque l'on traite des données provenant de sources multilingues. Cette étape permet notamment de filtrer les textes en fonction de la langue utilisée et d'appliquer des traitements adéquats. Les langues peu dotées ne sont malheureusement pas toujours supportées par les outils d'identification automatique. Dans cet article, nous présentons un outil d'identification automatique du créole guadeloupéen, qui repose sur une approche à base d'apprentissage automatique pour résoudre ce manque dans une certaine mesure. L'évaluation de notre modèle montre une précision de 89,96% sur l'identification de phrases en créole guadeloupéen provenant de différentes sources, et une précision de 91,04% sur l'identification de phrases dans un corpus de 103 langues.

ABSTRACT

Language Identification of Guadeloupean Creole

Language Identification is a useful preprocessing step that allows filtering and processing information on the best way possible, Improving the efficiency of Language Processing tasks. Under-resourced language, however, are often left out of most off-the-shelf applications for this task. In this article we present the Guadeloupean Creole Language Identification Tool, a Machine Learning (ML) approach to solve the lack of such applications for this under-resourced language. The evaluation of our model shows an 89.96% accuracy when classifying Guadeloupean Creole sentences from different sources and a 91.04% precision on the language when classifying sentences from 103 different languages.

MOTS-CLÉS : Identification Automatique de Langue, Créole Guadeloupéen, Apprentissage Automatique.

KEYWORDS: Language Identification, Guadeloupean Creole, Machine Learning.

1 Introduction

Language identification (LI) consist on determining in which language a given text has been written. As Jauhiainen et al. (2019) point out, Natural Language Processing (NLP) and Information Retrieval (IR) techniques generally assume that all documents given to a system are written in the same language. Because of this, being able to correctly identify and filter out documents that do not math the system's languages is a useful preprocessing step.

When you are guaranteed to have a variety of languages as input to your system, like with social network resources, this step becomes even more relevant. Cases like the SURICATE-Nat[1] project,

[1] See http://www.suricatenat.fr/Suricate-Nat/ for more information

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 53–58,

that collects and analyzes Twitter messages to facilitate information exchange from the ground during natural disasters, is a clear example of this. However, the LI task may not be as efficient in this kind of real life situations since most off-the-shelf tools do not support under-resourced languages, which in turn reduces the extent to which certain communities benefit from these projects.

Creole languages (languages that develop from the mix and simplification of other languages) are a good example of languages left out of most modern LI tools. Although these languages have been studied by linguists for centuries, many of them remain under-resourced when it comes to modern NLP applications. Their low number of speakers and a the lack of standardization are, no doubt, part of the reason behind that.

For this article we focused on one specific case that is usually not present on LI software: Guadeloupean Creole (GC, or Gwadloupéyen). This language developed from XVIII century French and a variety of West African languages (Delumeau, 2006) and accounts for around 600 000 speakers, from which close to 400 000 live in Guadeloupe (Colot and Ludwig, 2013) (a French archipelago in the Caribbean). For the most part it has been a spoken language and lacks a strong spelling standard, although there have been various attempts to change this (see Ludwig et al. (1990); Hazaël-Massieux (1993); Bernabé (2001) about the GEREC orthography). However, as Delumeau (2006) points out the language has become increasingly popular on Guadeloupe.

2 Related Work

Some LI off-the-shelf tools like LangDetect (Nakatani, 2010) and Compact Language Detector 2 (Sites, 2013) rely on probabilistic approaches, most specifically Naive Bayes classifiers. Some other methods, like Whatlang (Brown, 2013), rely on vector-space models. More recent methods are based on FastText (Armand et al., 2016) applying n-gram word embeddings and linear regression, like FastText's own language detection module (Joulin et al., 2016) and Whatthelang (Sangeeth, 2017).

All these methods have proven to be useful to solve LI tasks, but none of them provides support for Guadeloupean Creole. As we pointed out in the Introduction, under-resourced languages are usually excluded from modern NLP applications partially because of the lack of training data. For the case of GC, Millour and Fort (2018) mention some of the the existing resources of GC and detail the state of the art regarding GC corpora. However, the existing source are rather scarce and, to our knowledge, have not been applied to modern ML tasks.

3 Methodology

Following the most recent approaches we decided to use FastText as the base for our classifier. Not only do FastText-based models perform very well, but they are easy to train and to deploy. Our model consists of a FastText supervised classifier, which builds text representations by averaging n-grams and then performs multiple logistics regressions over this text representations. This model allows the use of subword features, which applies the n-grams at the character level and makes it possible to get information about the structure of words as well as supporting out of vocabulary (oov) words.

4 Dataset

As a baseline we decided to use the Tatoeba sentence dataset which contains more than 8 million sentences of 355 different languages. Most notably it, has 2 080 GC sentences. To reduce the training time and the noise added by languages without many samples, we limited the dataset to those languages with at least 1000 sentences, which left us with a total of 103 languages. Table 1 shows the distribution of the most common and uncommon languages on the dataset as well as that of GC.

Most common languages	Samples	Less common languages	Samples
English	1 319 616	Kapampangan	1 475
Russian	759 878	Cebuano	1 472
Italian	702 267	Ottoman Turkish	1 407
Turkish	685 782	Albanian	1 351
Esperanto	618 098	Picard	1 339
German	502 445	Khasi	1 320
French	425 191	Old East Slavic	1 307
Portuguese	358 570	Guarani	1 251
Spanish	317 954	Welsh	1 237
Hungarian	281 093	Slovenian	1 046

Table 1: Number of samples for the 10 most and less common languages in the tatoeba dataset.

To improve the models' ability to identify GC, we enriched the dataset with samples of the language from a diversity of sources. We used a set of transcriptions of spoken Guadeloupean Creole (Glaude, 2013), Caterina Bonan's Corpus of Guadeloupean Creole 2018, the 2012 Simenn Kréyòl collection of texts by the Academié de la Guadeloupe, the lyrics of two Guadeloupean Songs (the LKP song "gwada sé tan nou" and the Akiyo song "Jilo") and articles from the GC Wikipedia, which increased the number of GC sample sentences from 2080 to 4894.

All the datasets were preprocessed by lowercasing all the text and removing punctuation forms in each sentence. Then they were split in training and testing sets, with a 90% of the sentences going to the training set and the other 10% going to the testing set. Table 2 shows the number of GC samples present on each dataset.

Source	Training	Testing	Total
Tatoeba	1872	208	2080
Transcriptions	1354	150	1504
Caterina Bonan	689	78	767
Simmen Kréyòl	327	36	363
Chansons	100	11	111
Wikipedia	63	6	69
Total	4405	489	4894

Table 2: Number of Guadeloupean Creole samples on each dataset.

5 Experiment and Results

When instantiating the FastText supervised classifier we tried with 3 different sizes for the word vectors (16,32 and 64 dimensions) and enabled the use of subword features by setting up the char n-grams to a minimum size of 2 and a maximum size of 4. To set the other parameters of the model we tried the autotune method included in FastText, but after some tests we saw no significant improvement over the default settings, so we kept the default values on all the other parameters with the exception of the number of training epochs which showed an improvement when we doubled it from 5 to 10 epochs.

As mentioned above, the baseline was trained exclusively on the Tatoeba training set and another model was trained on the enriched GC dataset. To evaluate them, two tasks were defined: first we measured the accuracy of the models when presented only with the GC samples from each of the different sources, then we calculated the precision, recall and F1 score of the models when presented with samples from all the languages of the enriched dataset.

Tables 3 and 4 shows the result of the first and second task respectively. When classifying just GC sentences, as expected, the baseline performs well with the Tatoeba examples but accuracy drops with the other sources. In turn, the enriched model performs very well on all the sources. For the second task, when classifying samples from all 103 languages, the precision of the baseline is always slightly higher than that of the enriched model, but this comes at the cost of a significantly lower recall and F1 score.

On the first task, the 32 dimensions model performed better although the difference between all of them was very narrow. For the second task the 64 dimensions had better results that were also slightly more significant.

Source	Baseline (16)	Enriched (16)	Baseline (32)	Enriched (32)	Baseline (64)	Enriched (64)
Tatoeba	86.05%	87.98%	86.53%	88.94%	85.57%	88.94%
Transcriptions	57.33%	82.66%	50.00%	84.66%	52.66%	84.00%
Caterina Bonan	61.53%	94.87%	61.35%	94.87%	56.41%	94.87%
Simmen Kréyòl	66.66%	91.67%	58.33%	94.44%	63.88%	94.44%
Chansons	63.63%	90.90%	63.63%	90.90%	72.72%	90.90%
Wikipedia	83.33%	83.33%	83.33%	83.33%	83.33%	83.33%
All datasets	71.37%	87.73%	68.87%	88.95%	68.91%	88.75%

Table 3: Accuracy of each model when classifying Guadeloupean Creole sentences from each of the sources

	Baseline (16)	Enriched (16)	Baseline (32)	Enriched (32)	Baseline (64)	Enriched (64)
Precision	96.14%	94.07%	96.27%	91.04%	96.28%	95.59%
Recall	71,37%	87.73%	68.87%	88.95%	68.91%	88.75%
F1 score	81.92%	90.79%	80.19%	89.96%	80.33%	92.04%

Table 4: Precision, recall and F1 score for Guadeloupean Creole of each model when classifying sentences from 103 languages from the enriched dataset

6 Error Analysis

We extracted the most common errors made by the system. Table 5 shows which languages were most commonly assigned to real GC samples. As expected, French was the most common misclassification in all but one models, in which it was the second most common. There is no clear second nor third place but it is worth noticing the constant appearance of Austronesian languages. On the other hand, Table 6 shows which languages were most commonly misclassified as CG. In this case there is no clear first place, although Spanish seems to be the most consistent error, however the distribution among the misclassified languages seems to be more uniform.

Model	Misclassification	Error	Misclassification	Error	Misclassification	Error
Baseline(16)	Kapampangan	16%	French	14%	Tagalog	6%
Enriched (16)	French	20%	Spanish	10%	Waray	8%
Baseline (32)	French	18%	Kapampangan	9%	Irish	5%
Enriched (32)	French	24%	Waray	11%	Esperanto	11%
Baseline (64)	French	18%	Kapampangan	14%	Interlingue	5%
Enriched (64)	French	22%	Esperanto	9%	Spanish	7%

Table 5: Three most common errors when classifying Guadeloupean Creole as another language and how much of the total number of errors each of them represents

Model	Real Language	Error	Real Language	Error	Real Language	Error
Baseline(16)	Spanish	14%	Low Saxon	7%	French	7%
Enriched (16)	Spanish	15%	French	11%	Finnish	11%
Baseline (32)	Breton	15%	Cebuano	15%	Lojban	15%
Enriched (32)	Kotava	14%	Spanish	12%	Portugues	12%
Baseline (64)	Spanish	23%	Lojban	15%	Interlingue	15%
Enriched (64)	Lojban	10%	Spanish	10%	Esperanto	10%

Table 6: Three most common errors when classifying another language as Guadeloupean Creole and how much of the total number of errors each of them represents

7 Conclusion

We produced a reliable Language Identification tool for Guadeloupean Creole and proved that enriching the training of the model with few but diverse sources of the language helps to improve the performance. We also produced a public version of our models along with a python wrapper and a terminal tool for ease of use[2]. There is still much room for improvement both in the design and processing of a richer set of GC examples as well as on the settings of the classification model. We hope that this article will motivate other researchers to work on the implementation of NLP tools of under-resourced languages.

[2]See `https://gitlab.com/williamsotomartinez/gclit/`

References

Armand, J., Grave, E., Bojanowski, P., and Tomas, M. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Bernabé, J. (2001). *La graphie créole*. Ibis rouge.

Bonan, C. (2018). Online corpus of guadeoupean creole. `https://caterinabonan.com/corpus-of-guadeloupean-creole/`Visited on 18/06/2020.

Brown, R. D. (2013). Selecting and weighting n-grams to identify 1100 languages. In *International Conference on Text, Speech and Dialogue*, pages 475–483. Springer.

Colot, S. and Ludwig, R. (2013). Guadeloupean and martinican creole. In Michaelis, S. M., Maurer, P., Haspemath, M., Huber, M., and Revis, M., editors, *The Survey of Pidgin and Creole Languages: Volume 2*. Oxford University Press.

Delumeau, F. (2006). *Une description linguistique du Creole Guadeloupéen dans la perspective de la genération automatique d'enonces.* PhD thesis, Université de Nanterre - Paris.

Glaude, H. (2013). Corpus Créoloral. oai:crdo.vjf.cnrs.fr:crdo-GCF, SFL Université Paris 8 - LLL Université Orléans.

Hazaël-Massieux, M.-C. (1993). Ecrire en créole(oralité et écriture aux antilles). *Journal of French Language Studies*.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Ludwig, R., Montbrand, D., Poullet, H., and Telchid, S. (1990). Abrégé de grammaire du créole guadeloupéen. *Dictionnaire créole français (Guadeloupe), avec un abrégé de grammaire créole et un lexique françaiscréole*, pages 17–38.

Millour, A. and Fort, K. (2018). Krik: First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen. In *CCURL 2018* , Miyazaki, Japan.

Nakatani, S. (2010). Language detection library for java. Software available at `http://code.google.com/p/language-detection/` (last updated on March 2014).

Sangeeth, K. (2017). Whatthelang. Software available at `https://github.com/indix/whatthelang`.

Sites, D. (2013). Compact language detector 2. Software available at `https://github.com/CLD2Owners/cld2` (last updated on August 2015).

Lexical encoding of multiword expressions in XMG

Agata Savary[1] Simon Petitjean[2] Timm Lichte[3]
Laura Kallmeyer[2] Jakub Waszczuk[2]
(1) University of Tours, France
(2) Heinrich Heine Universität Düsseldorf, Germany
(3) University of Tübingen, Germany
first.last@univ-tours.fr
last@phil.uni-duesseldorf.de
first.last@uni-tuebingen.de

ABSTRACT

Multiword expressions (MWEs) exhibit both regular and idiosyncratic properties. Their idiosyncrasy requires lexical encoding in parallel with their component words. Their (at times intricate) regularity, on the other hand, calls for means of flexible factorization to avoid redundant descriptions of shared properties. However, so far, non-redundant general-purpose lexical encoding of MWEs has not received a satisfactory solution. We offer a proof of concept that this challenge might be effectively addressed within eXtensible MetaGrammar (XMG), an object-oriented metagrammar framework. We first make an existing metagrammatical resource, the FrenchTAG grammar, MWE-aware. We then evaluate the factorization gain during incremental implementation with XMG on a dataset extracted from an MWE-annotated reference corpus. This paper is part of a larger publication to appear (Savary et al., 2020).

RÉSUMÉ

Codage lexical d'expressions polylexicales en XMG

Les Expressions polylexicales (EP) possèdent des propriétés à la fois régulières et idiosyncratiques. Leur idiosyncrasie requiert un codage lexical au même titre que celui des mots qui les composent. D'autre part, leur régularité (parfois complexe) nécessite des moyens de factorisation afin d'éviter des descriptions redondantes des propriétés partagées. À ce jour, il n'existe pas de solution idéale pour le codage lexical généraliste et non redondant des EP. Dans cet article nous présentons une preuve de concept que ce défi pourrait être relevé dans le cadre de XMG (eXtensible MetaGrammar), qui est un formalisme métagrammatical orienté-objet. Nous montrons comment une ressource métagrammaticale existante, FrenchTAG, peut être étendue pour couvrir les EP. Nous évaluons le gain en terme de factorisation de cette ressource lors de son développement incrémental. Cette expérience est menée sur un jeu de données extrait d'un corpus de référence annoté en EP. Cet article est extrait d'un publication plus large à venir (Savary et al., 2020).

MOTS-CLÉS : expressions polylexicales, métagrammaire, XMG.

KEYWORDS: multiword expressions, metagrammar, XMG.

Multiword expressions (MWEs) are combinations of words which encompass heterogeneous linguistic objects such as idioms (IDs : *to* ***pull*** *one's* ***leg***), compounds (*a* ***hot dog***), light verb constructions (LVCs : *to* ***pay*** *a* ***visit***), inherently reflexive verbs (IRVs : ***s'apercevoir*** 'perceive oneself'⇒'realize' in French), rhetorical figures (***as busy as a bee***), or named entities (*the* ***Sea of Tranquility***). Their

most pervasive and challenging feature is their non-compositional semantics, i.e. the fact that their meaning cannot be deduced from the literal meanings of their components, and from their syntactic structures, in a way deemed regular for the given language. For this reason, as well as because of their pervasiveness in texts, MWEs constitute a major challenge in semantically oriented NLP applications.

But MWEs also exhibit unexpected behavior on other levels of linguistic analysis including the lexical, morphological and syntactic ones. These properties can be *defective* or *restrictive* (Lichte et al., 2019). A defective property excludes a literal interpretation of the MWE, e.g. (EN) *a* ***lesser yellowlegs*** 'a shore bird species' cannot be understood literally because of the lack of number agreement between the determiner and the head noun. A restrictive property reduces the number of possible surface realizations of the MWE with respect to the literal reading. For instance in example (3), the possessive determiner has to agree with the subject, otherwise the expression can only be understood literally as in *#John crossed her fingers.* [1] Since defective and restrictive properties help distinguish literal from idiomatic readings of MWEs, their description and processing are important both for linguistic modeling and for NLP applications, including MWE identification (Constant et al., 2017).

When characterizing MWEs, some authors (Grégoire, 2010; Przepiórkowski et al., 2014) oppose the *regular* behavior of "free" phrases (i.e. those obeying the rules of a "regular" grammar), like (1), to the *idiosyncratic* behavior of MWEs, like (2)–(4).

(1) *John broke my mug*

(2) *John* ***broke*** *his/our* ***fall*** 'John made his/our fall less forceful'

(3) *John* ***crossed*** *his* ***fingers*** 'John hoped for good luck'

(4) *John* ***held*** *his* ***tongue*** 'John refrained from expressing his view'

Some others point out that regularity is a matter of scale rather than a binary phenomenon (Gross, 1988; Herzig Sheinfux et al., 2015). We take the latter stand, and extend it by assuming that the degree of regularity is a feature of linguistic properties on the one hand, and of MWEs on the other hand (Lichte et al., 2019). Firstly, the more (resp. less) objects share a certain property, the more it is regular (resp. idiosyncratic). For instance, allowing a possessive determiner in a Verb-Det-Noun construction is more regular than imposing that it agrees with the subject, because the former applies to (1)–(4), while the latter is limited to (3)–(4). Still the latter is not fully irregular since it is shared by many expressions. Secondly, in (3), while the direct object of the verb *to cross* is lexicalized (has to be realized by the lexeme *finger*), the subject is not. While the noun does not admit adjectival modifiers (*#He crossed his long fingers.*), passivization is allowed (***fingers crossed***). While the noun has to occur in plural, the verb can be inflected freely, etc. Thus, this MWE combines more regular properties (e.g. a free subject) with more idiosyncratic ones (e.g. a lexically and morphologically fixed object). Also, the MWE in (4) has the same properties (with the number of the noun fixed to singular instead of plural) except that passivization is not allowed (*#His tongue was held*). Therefore, the degree of regularity of the MWE in (3) can be considered higher than of the one in (4).

Because MWEs exhibit (more or less) idiosyncratic properties, their modeling has to include lexical encoding, i.e. MWEs should become separate lexical entries, additionally to their single-word components. The main challenge is then to account for the irregularity of a MWE, while avoiding redundancy, i.e. repeated description of common properties. For instance, the subject-possessive agreement is shared by (3)–(4) and many other MWEs, so its formalization should preferably be

1. The hash symbol # signals the loss of the idiomatic reading. Lexicalized components of a MWE, i.e. those always realized by the same lexemes, are marked in boldface.

done only once, rather than repeatedly for each MWE lexicon entry. Our state-of-the-art studies have shown that no previous work seems to have addressed this challenge in a satisfactory way.

In this work, we aim at providing a proof of concept that non-redundant lexical encoding of MWEs can be effectively achieved in an object-oriented metagrammar-based approach. We use XMG (Crabbé et al., 2013; Petitjean et al., 2016), a declarative constraint-based description language in which more or less regular tree structures are modeled via a hierarchy of classes. Higher (more abstract) classes encode more elementary and less constrained structures. Lower (more specific) classes combine higher ones and impose new constraints on these combinations. Both single-word lexemes and MWEs are then expressed as lexical entries assigned to particular low-level classes (usually leaves) of this class hierarchy. The description is independent of a particular grammatical framework but XMG comes with metagrammar compilers into several formalisms including Tree Adjoining Grammar (TAG). We therefore test our proposal on FrenchTAG (Crabbé, 2005), a pre-existing XMG resource which implements a large fragment of a reference grammar of French (Abeillé, 2002). We show how FrenchTAG can be adapted and extended so as to accommodate a small subset of verbal MWEs (VMWEs) of different syntactic structures and of varying degrees of syntactic flexibility. We evaluated the proposal on a dataset based on the PARSEME corpus of VMWEs (Savary et al., 2018). The experiment showed that adding MWE descriptions to a general grammar can be done elegantly by introducing interface constraints in pre-existing classes (to account for restrictive properties), and by adding some new classes (to account for defective properties and for various syntactic structures of lexicalized verbal arguments).

This abstract is part of a larger publication to appear (Savary et al., 2020).

Références

Abeillé, A. (2002). *Une grammaire électronique du français*. CNRS Editions.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword Expression Processing : A Survey. *Computational Linguistics*, 43(4) :837–892.

Crabbé, B., Duchier, D., Gardent, C., Roux, J. L., and Parmentier, Y. (2013). XMG : extensible metagrammar. *Computational Linguistics*, 39(3) :591–629.

Crabbé, B. (2005). *Représentation informatique de grammaires d'arbres fortement lexicalisées : le cas de la grammaire d'arbres adjoints*. PhD thesis, Université Nancy 2.

Grégoire, N. (2010). DuELME : a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2) :23–39.

Gross, G. (1988). Degré de figement des noms composés. *Langages*, 90 :57–71. Paris : Larousse.

Herzig Sheinfux, L., Arad Greshler, T., Melnik, N., and Wintner, S. (2015). Hebrew verbal multiword expressions. In Müller, S., editor, *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, pages 122–135, Stanford, CA. CSLI Publications.

Lichte, T., Petitjean, S., Savary, A., and Waszczuk, J. (2019). Lexical encoding formats for multiword expressions : The challenge of "irregular" regularities. In Parmentier, Y. and Waszczuk, J., editors, *Representation and parsing of multiword expressions : Current trends*, pages 1–33. Language Science Press, Berlin.

Petitjean, S., Duchier, D., and Parmentier, Y. (2016). XMG 2 : Describing description languages. In Amblard, M., de Groote, P., Pogodalla, S., and Retoré, C., editors, *Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996-2016) - 9th International Conference, LACL 2016, Nancy, France, December 5-7, 2016, Proceedings*, volume 10054 of *Lecture Notes in Computer Science*, pages 255–272.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., and Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V., editors, *Multiword expressions at length and in depth : Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Savary, A., Petitjean, S., Lichte, T., Kallmeyer, L., and Waszczuk, J. (2020+). Object-oriented lexical encoding of multiword expressions : Short and sweet. *Lexique*, forthcoming.

Enregistrements de longue durée: Opportunités et défis

Lucas Gautheron[1] Marvin Lavechin[1, 2] Rachid Riad[1, 2, 3] Camila Scaff [4,1]
Alejandrina Cristia[1]
(1) Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
(2) CoML Team, INRIA, Paris, France
(3) Laboratoire de Neuropsychologie Interventionnelle, Département d'Etudes cognitives, ENS, INSERM, UPEC, PSL University, Paris, France
(4) University of Zurich, Zurich, Switzerland
alecristia@gmail.com*

RÉSUMÉ

Bénéficiant d'améliorations technologiques récentes, des appareils audio légers et portables sont désormais capables d'enregistrer des dizaines d'heures sans interruption. Nous proposons une description générale de cette technique pour l'étude de la parole, et faisons le point sur ses avantages et inconvénients. Grâce à elle, les linguistes de terrain bénéficient d'un accès unique au langage dans un contexte plus naturel. Cependant, ces enregistrements restent difficiles à annoter manuellement ou automatiquement, en raison de leur durée, du bruit, et de la sensibilité des informations qu'ils peuvent contenir. Des outils *open-source* plus facilement appopriables, auxquels les spécialistes des technologies de la parole peuvent contribuer, favorisent la reproductibilité des travaux des chercheurs. En outre, de nouvelles approches aux techniques d'annotation manuelles ou automatiques rendent cette technique opérationnelle et prometteuse.

ABSTRACT

Longform recordings : Opportunities and challenges

Technological developments have allowed the development of lightweight, wearable recorders that collect audio (including speech) lasting up to a whole day. We provide a general description of the technique and lay out the advantages and drawbacks when using this methodology. Field linguists may gain a uniquely naturalistic viewpoint of language use as people go about their everyday activities. However, due to their duration, noisiness, and likelihood of containing sensitive information, longform recordings remain difficult to annotate manually. Open-source tools improve reproducibility and ease-of-use for researchers, to which end speech technologists can contribute. Additionally, new approaches to human and automated annotation make the study of speech in longform recordings increasingly feasible and promising.

MOTS-CLÉS : enregistrements longs ; validité écologique ; traitement automatique de la parole.

KEYWORDS: daylong recordings ; ecological validity ; automatic speech processing.

Recent years have seen the rise of data collection through wearable, light-weight and unobtrusive devices that collect audio for tens of hours at a time, allowing a uniquely naturalistic viewpoint of language use as people go about their everyday activities. Over nearly a decade, our team gained first-hand experience with the incredible benefits as well as the painful points of this data collection

∗. We acknowledge ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017; and the J. S. McDonnell Foundation

(a)

(b)

(c)

FIGURE 1 – Examples of wearable recorders. (a) Tsimané child wearing a LENA device in the front pocket of a purpose-made vest. (b) Smart watch recording audio, heart rate, and movement, adapted from Fig 1 in (Liaqat et al., 2018). (c) Body camera on a South Carolina police officer (Ryan Johnson, CC BY-SA 2.0).

technique. By now, our lab has over 20,000 hours of audio, capturing language experiences of over 1,000 children, learning one or more of 16 typologically diverse languages. We provide a brief introduction to this technique, in the hope of allowing our colleagues to decide when it may be a useful tool to add to their kit (for detailed information, see (Casillas et al., 2019)).

1 Interest of the method

1.1 Providing decisive evidence on long-standing debates

Short recordings and controlled data elicitation provide crucial information about language perception and production, but we still know little about spontaneous language use in naturalistic environments. A new window on this was opened by daylong recordings. The technique has already proven fruitful in the field of language acquisition, from where we provide several examples.

One of the key theoretical questions in the field was whether language development is mainly driven by infants wish to communicate, or other processes. Kim Oller and colleagues have been studying development of speech in infancy for many decades, and had discovered that there are speech-like sounds, called protophones, even in young infants – but this still did not settle the question of *why* infants vocalize like this. Only recently, using infant-centered daylong recordings, Oller and colleagues were able to show that, in 6-month-olds, vocalizations were more advanced when the

infant was *not* being talked to, suggesting these vocalizations are endogeneous rather socially driven (Lee et al., 2018). Other results also contradicted prior beliefs that cries were more abundant than protophones at early stages, as the opposite was true even among preterm and fullterm newborns, suggesting in a way that infants are literally born to produce speech (Oller et al., 2019).

The interest for this technique is increasing beyond the language acquisition community. Additional applications being explored include the relationship between social interaction and well-being (Sun et al., 2019), activities among adults suffering from pulmonary diseases (Wu et al., 2018), and measurement of speech and language correlates of neurodegenerative diseases (Riad et al., 2020).

1.2 Drawbacks and challenges

The technique of long-form recordings also comes with its own challenges and limitations. Extracting information from the data may be challenging. The technique produces large amounts of audio which cannot be manually annotated as a whole. Automated tools are thus often required to extract sections and/or annotate the data automatically. However, there are no off-the-shelf solutions for automated annotation, which instead require active development by experts in speech and language technology. Still, even when these tools are developed, they do not have perfect accuracy, and thus it may be impossible to detect small effects. Indeed, contrary to lab-controlled experiments, the data suffer from significant background noise, and are potentially subject to a variety of soundscapes.

Furthermore, recordings might contain confidential or sensitive information, including from people who are accidentally recorded and have therefore not provided informed consent. As a result, researchers often need to ponder difficult ethical and legal questions, for which they may need advice from experts in law and ethics, who may not be familiar with long-form recordings from wearables. Storage on embedded devices or on untrustworthy third-party services such as cloud platforms might require encryption. Data transfers should be secure to prevent leaks beyond the research community.

1.3 When should LIFT members consider using and/or contributing to this technique ?

As with any (new) technique, one should make sure it is appropriate for its research purposes before engaging in it. For readers who are considering **collecting data** with long-form recordings from wearable, we clarify that these recordings are most valuable when (1) ecological validity is key (where e.g. elicitation is inappropriate), (2) unbiased sampling is important, and (3) the phenomenon studied occurs frequently in language use and it is robust to the presence of ambient noise (particularly if automated annotation will be used). We provide some information specifically for LIFT members here (see Casillas et al., 2019 for more examples).

We expect that field linguists in LIFT will find it particularly useful to collect data with wearables when interested in language use in situations where their presence as an observer may not allow a behavior to develop naturally, and when their informants find it hard to report on the use of a form (or their reports may not reflect actual use). To give a specific example, one of our collaborators uses the samples to study patterns of language switching in a highly multilingual community. After her informant has worn the recording device for a whole day, she extracts sections with speech randomly throughout the day, and uses them as prompts to discuss with her informants which language was

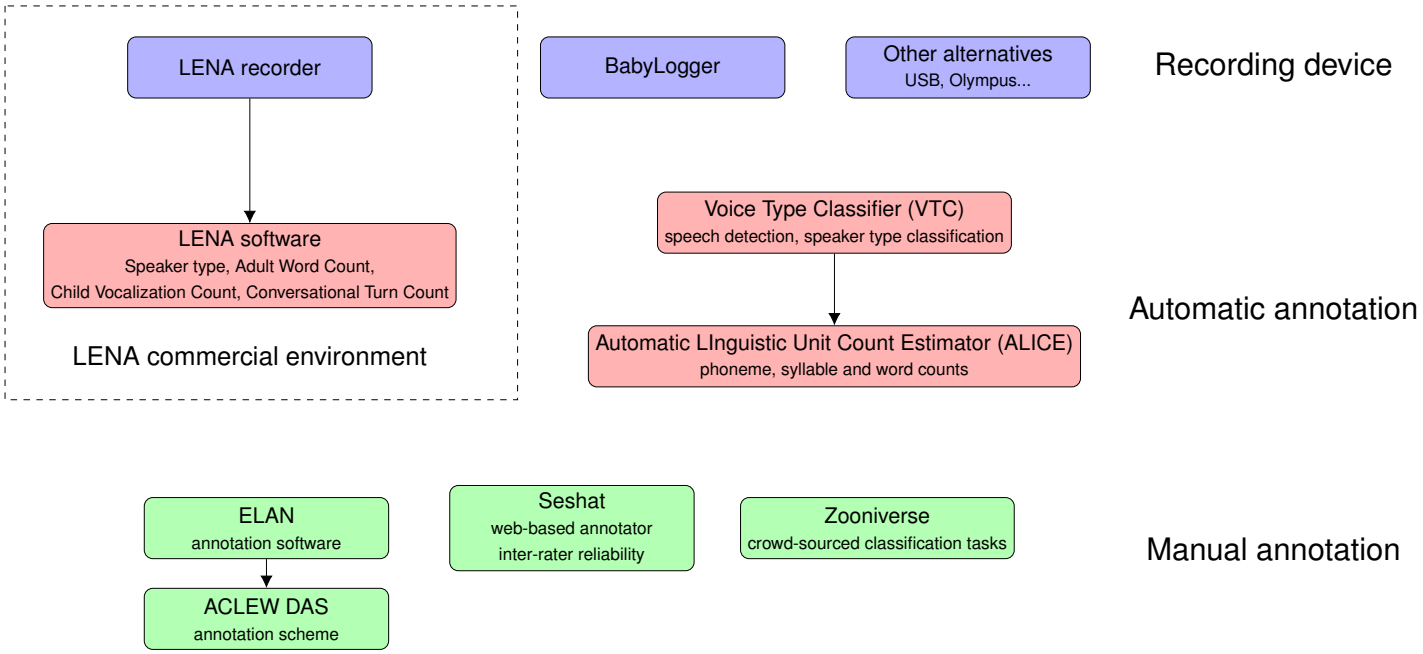

FIGURE 2 – Overview of current solutions for longform recordings, mainly applied to early language acquisition research.

being used, by whom, and why they think that language was used rather than the others (e.g., because of who was overhearing the conversation, or because of the topic).

As for **contributing to the technique's development**, computational linguists in LIFT will find it easiest to contribute to this literature if they already have some experience working with speech, as it has so far proved difficult to use automatic speech recognition (ASR) to generate automated transcriptions. Even a recent study on typical English-speaking adults had humans transcribe the clips, rather than using ASR. But we hope this will not discourage LIFT computational linguists who are interested in this emergent field, as there are many opportunities to start dipping one's toes in, for instance via participation in public challenges (Ryant et al., 2019; Schuller et al., 2019).

2 Tools and ecosystem

2.1 LENA

The most commonly used hardware and software for automatically analysing infant's speech is the Language ENvironment Analysis (LENA©), a commercial lightweight recorder worn in a specially designed vest associated with a closed-source speech processing algorithm. A key strength of LENA is that it provides users with an end-to-end pipeline to collect and analyze daylong recordings, making it efficient and easy to use. LENA's recorders have been designed to be unobtrusive and easy-to-wear to improve the ecological validity of field observations by reducing observer bias. The hardware can record up to 16 hours and can only be analysed by the associated software. Their software was trained on American English input to children aged 0-4 years. It generates an automatic analysis of three key estimates of the child's environment : the number of words spoken by a nearby adult ; the number of times the child made any kind of linguistically relevant vocalization (i.e., speech or babble and

excluding vegetative noises, laughter or crying) ; and the number of exchanges between an adult and the child within a five seconds window, considered as "turns". The LENA technology has been used in numerous studies (Dykstra Steinbrenner et al., 2012; Vandam et al., 2015; Ferjan Ramírez et al., 2020).

The LENA system has been found to be fairly accurate in quantifying children's language environment. A recent study comparing the algorithm automated measures to manual human transcriptions found high correlations for the two measures quantifying the number of adult and child's vocalizations and a moderate correlation for the number of turns (Cristia et al., 2020). Studies on annotations automatically extracted by the LENA speech processing pipeline did not reveal great differences in adult word count accuracy between American English and other languages including Swedish (Schwarz et al., 2017) and French (Canault et al., 2015).

However, there have also been some reports of a problematic level of performance for some of LENA's outputs (notably the key child's vocalization recall) (Cristia et al., 2019). This means researchers using LENA output are relying on a noisy analysis, which potentially hides small statistical effects.

Moreover, despite its use in many linguistic studies and its wide acceptation in the child language community, LENA imposes several limiting factors to scientific progress. There is currently no way to build upon LENA speech processing models as their software is closed source, and gathering information about design choices and their potential impact on performance remains a tedious task. Concerning the recorder, LENA designers' hardware choices cannot be revised. This has raised multiple questions that remain to be answered, mainly : Does a single-channel microphone allow us to capture the full complexity of the child's language environment ? Can alternative models provide us with more faithful metrics of this environment ? Addressing these questions must start by creating open-source alternatives to LENA.

	Autonomy		**Audio properties**				
Device	Battery	Storage	Channels	Sampling rate	Bit depth	Weight	Cost (US$)
LENA	30 h	15 h	1	16 kHz[1]	16	200 g	300
BabyLogger	24 h	SD[2]	4	16 kHz	16	200 g	500
USB	15 h	150 h	1	16 kHz	8	50 g	20
Olympus	25 h	SD[2]	1	22 kHz[3]	32[3]	400 g	300

TABLE 1 – Technical characteristics of various recording devices suitable for child-centered audio collection.

2.2 Alternative tools

We have been leading an effort to build open-source alternatives to LENA speech processing algorithms, providing researchers with models that have similar outputs to the ones returned by LENA, as well as undertaking systematic comparisons of these models with their LENA counterpart. We released a voice type classifier (Lavechin et al., 2020) classifying audio segments into vocalizations

1. Audio undergoes a 10 kHz low-pass filter.
2. Limited by the mini SD card the user fits in.
3. Can be adjusted by the user.

produced by the child wearing the recording device, vocalizations produced by other children, adult male speech, and adult female speech. Building upon this effort, a linguistic unit count estimator (Räsänen et al., 2020) has been developed, allowing users to count the number of words, syllables or phonemes produced by adult speakers. These two models have been shown to outperform their LENA counterpart. We redirect our readers to those papers for more in-depth analysis.

As for the hardware, there exist multiple lightweight recording devices available in the market that one might use to acquire speech, from body mounted cameras to digital voice recorders, each with their own hardware specification (see Fig. 1). There are fewer alternatives to specifically acquire child's speech (Table 1) as these devices require particular safety norms and design to be wearable by young children. One interesting alternative that has been specifically designed for child data acquisition is the BabyLogger (Cao et al., 2018), using an array of four microphones, as opposed to one for LENA. The BabyLogger also performs on-the-fly encryption, protecting the privacy of the participants in case the device is lost or stolen. In the context of patient monitoring, smartwatches (1b) paired with smartphones have also been employed, allowing teletransmission of the data to a remote server at the expense of lesser audio quality (because of bandwidth limitations) and lower duty cycles (to avoid premature battery shortage) (Liaqat et al., 2018). However, more work is needed to know whether or not different hardware specifications might lead to different views on language environments.

2.3 Manual annotations

Because of the noisy nature of the recordings, today's classification algorithms might perform too poorly for practical analyses. Even low-level tasks such as speech detection or diarization can be hard to achieve automatically (Ryant et al., 2019).

Therefore, to evaluate these algorithms for specific corpora and for improving these machine learning models, additional manual annotations are required. In these difficult audio data, human annotations take about 40 times the audio duration. Typical datasets of daylong child recordings can contain thousands of hours of audio, and would require hundreds of thousands of work time to be fully annotated. Nonetheless, it is possible to reduce the amount of audio to annotate manually in a few ways, including by performing random sampling of the data with uniform or non-uniform priors.

We developed a manual annotation scheme to help researchers annotating daylong recordings in a systematic way, thus improving reproducibility and comparisons across studies (Casillas et al., 2017). In a nutshell, our annotation scheme allows researchers to both contribute to machine learning efforts and serve their research goals : Talkers are segmented, and certain layers of information can be added optionally, including transcription or classification into fixed classes (e.g., vocalization type : crying, laughing, canonical, non-canonical).

When adding layers of annotations on top of audio data, researchers face many challenges to handle their campaign of annotations : the *problem around files management* (ex : character-encoding problems, incorrect naming of files), the *non-conformity of the annotations* to the schema established by researchers (misuse of symbols), and the *inconsistency of the annotations* (not properly annotated). That is why we introduced the Seshat software (Titeux et al., 2020). It allows researchers to easily customise and standardize annotations and manage annotators. Finally, to measure how "reliable" are the annotations, we implemented an open-source version of the Gamma Agreement measure in Python (Titeux and Riad, 2020). This allows to measure inter and intra annotator agreement for the type of annotations around speech data.

It would be impractical to rely solely on experts to manually annotate such volumes of audio. Most recently, our team launched a crowd-sourcing project on Zooniverse asking citizens' help to solve simple classification tasks on short audio chunks drawn from the daylong recordings, which proved quite accurate and will allow data annotation at a much larger scale (Semenzin et al., 2020).

3 Conclusion

Despite the many challenges that data from wearables bring, we believe this is a technique fitting to the 21st century, and merits our colleagues' attention as a potential tool in their kit. We highly recommend it to those who are particularly concerned by ecological validity of their observations, and who are interested in phenomena that is common and can be studied from surface (acoustic) features. This is a field in expansion, with at least one speech technology challenge on average over the last 3 years, which is ideal for promoting interactions between speech technologists and field linguists.

Références

Canault, M., Normand, M.-T. L., Foudil, S., Loundon, N., and Thai-Van, H. (2015). Reliability of the language ENvironment analysis system (LENA™) in european french. *Behavior Research Methods*, 48(3) :1109–1124.

Cao, X.-N., Dakhlia, C., Del Carmen, P., Jaouani, M.-A., Ould-Arbi, M., and Dupoux, E. (2018). Baby Cloud, a technological platform for parents and researchers. In *LREC 2018 - 11th edition of the Language Resources and Evaluation Conference*, Proceedings of LREC 2018, Miyazaki, Japan.

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., and Sloetjes, H. (2017). A new workflow for semi-automatized annotations : Tests with long-form naturalistic recordings of childrens language environments. In *Proc. Interspeech 2017*, pages 2098–2102.

Casillas, M., Cristia, A., Zwaan, R., and Dingemanse, M. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra : Psychology*, 5(1). 24.

Cristia, A., Bulgarelli, F., and Bergelson, E. (2020). Accuracy of the language environment analysis system segmentation and metrics : A systematic review. *Journal of Speech, Language, and Hearing Research*, 63(4) :1093–1105.

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., and Bergelson, E. (2019). A thorough evaluation of the language environment analysis (lena) system. *Behavior Research Methods*.

Dykstra Steinbrenner, J., Sabatos-DeVito, M., Irvin, D., Boyd, B., Hume, K., and Odom, S. (2012). Using the language environment analysis (lena) system in preschool classrooms with children with autism spectrum disorders. *Autism : the international journal of research and practice*, 17.

Ferjan Ramírez, N., Lytle, S. R., and Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, 117(7).

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., and Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *Interspeech*.

Lee, C.-C., Jhang, Y., Relyea, G., Chen, L.-m., and Oller, D. K. (2018). Babbling development as seen in canonical babbling ratios : A naturalistic evaluation of all-day recordings. *Infant Behavior and Development*, 50 :140–153.

Liaqat, D., Wu, R., Gershon, A., Alshaer, H., Rudzicz, F., and de Lara, E. (2018). Challenges with real-world smartwatch based audio monitoring. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, WearSys '18, page 54–59, New York, NY, USA. Association for Computing Machinery.

Oller, D. K., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Lee, C.-C., Bowman, D. D., Long, H. L., Buder, E. H., and Vohr, B. (2019). Preterm and full term infant vocalization and the origin of language. *Scientific Reports*, 9(1) :14734.

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., and Casillas, M. (2020). Alice : An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*, pages 1–18.

Riad, R., Titeux, H., Lemoine, L., Montillot, J., Bagnou, J. H., Cao, X. N., Dupoux, E., and Bachoud-Lévi, A.-C. (2020). Vocal markers from sustained phonation in huntington's disease. *Interspeech*.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019). The second dihard diarization challenge : Dataset, task, and baselines. *arXiv preprint arXiv :1906.07839*.

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Vollmann, R., Roelen, S.-D., Schnieder, S., Bergelson, E., et al. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech*, pages 2378–2382.

Schwarz, I.-C., Botros, N., Lord, A., Marcusson, A., Tidelius, H., and Marklund, E. (2017). The LENA system applied to swedish : Reliability of the adult word count estimate. In *Interspeech 2017*. ISCA.

Semenzin, C., Hamrick, L., Seidl, A., Kelleher, B., and Cristia, A. (2020). Towards large-scale data annotation of audio from wearables : Validating zooniverse annotations of infant vocalization types. (Accessed on 11/25/2020).

Sun, J., Harris, K., and Vazire, S. (2019). Is well-being associated with the quantity and quality of social interactions ? *Journal of Personality and Social Psychology*.

Titeux, H. and Riad, R. (2020). *pygamma-agreement : Gamma γ measure for inter/intra-annotator agreement in Python*.

Titeux, H., Riad, R., Cao, X.-N., Hamilakis, N., Madden, K., Cristia, A., Bachoud-Lévi, A.-C., and Dupoux, E. (2020). Seshat : A tool for managing and verifying annotation campaigns of audio data. In *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France.

Vandam, M., Oller, D. K., Ambrose, S., Gray, S., Richards, J., Gilkerson, J., Silbert, N., and Moeller, M. (2015). Automated vocal analysis of children with hearing loss and their typical and atypical peers. *Ear and hearing*, 36.

Wu, R., Liaqat, D., de Lara, E., Son, T., Rudzicz, F., Alshaer, H., Abed-Esfahani, P., and Gershon, A. S. (2018). Feasibility of using a smartwatch to intensively monitor patients with chronic obstructive pulmonary disease : Prospective cohort study. *JMIR mHealth and uHealth*, 6(6) :e10046.

Modèles d'annotations morphologiques pour le traitement de données multivariées de l'arménien

Chahan Vidal-Gorène[1] Victoria Khurshudyan[2] Anaïd Donabédian[2]
(1) École Nationale des Chartes-PSL, 65 rue Richelieu, 75002 Paris, France
(2) SeDyL, UMR8202, INALCO, CNRS, IRD, 65 rue des Grands Moulins, 75013 Paris, France
chahan.vidal-gorene@chartes.psl.eu, victoria.khurshudyan@inalco.fr,
anaid.donabedian@inalco.fr

RÉSUMÉ

L'arménien est une langue comprenant de multiples variantes très inégales en termes de ressources disponibles en TAL. Nous avons entraîné un RNN pour réaliser l'annotation morphologique de différentes variantes de l'arménien, afin d'en comparer les résultats avec une approche par règles. Plusieurs tests ont permis d'évaluer la réutilisation d'un modèle non spécialisé de lemmatisation et de POS-tagging pour des variétés linguistiques sous représentées. Notre recherche s'est concentrée sur trois dialectes et a été étendue à l'arménien occidental, avec une précision moyenne de 94,00% en lemmatisation et 97,02% en POS-tagging, ainsi que sur une éventuelle réutilisation des modèles pour couvrir différentes autres variétés de l'arménien (jusqu'à 81% en POS-tagging). Nous montrons qu'une approche par RNN peut être une alternative valable à une approche par règles dans le cas d'une langue peu dotée et multivariées, en tenant compte de facteurs tels que la rapidité de traitement, la réutilisabilité pour différentes variétés d'une langue, et du gain qualitatif significatif en annotation morphologique.

ABSTRACT

Morphological Annotation Models for Armenian Multivariational Data Processing

Armenian is a language with significant variation and unevenly distributed NLP resources for different varieties. An attempt is made to process an RNN model for morphological annotation on the basis of different Armenian data and to compare the annotation results of RNN and rule-based models. Different tests were carried out to evaluate the reuse of an unspecialized model of lemmatization and POS-tagging for under-resourced language varieties. The research focused on three dialects and further extended to Western Armenian with a mean accuracy of 94,00% in lemmatization and 97,02% in POS-tagging, as well as a possible reusability of models to cover different other Armenian varieties (up to 81% in POS-tagging). It is argued that an RNN-based model can be a valid alternative to a rule-based one giving consideration to such factors as time-consumption, reusability for different varieties of a target language and significant qualitative results in morphological annotation results.

MOTS-CLÉS : variation linguistique, linguistique de corpus, arménien, RNN, règles-dictionnaire, étiquetage morphologique.

KEYWORDS: linguistic variation, corpus linguistics, Armenian, RNN, rule-based, morphological tagging.

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 71–81,

1 Introduction

Le renouvellement des ressources du traitement automatique des langues (TAL) au moyen de diverses ressources d'intelligence artificielle et réseaux de neurones (comme les réseaux de neurones récurrents, ci-après RNN) ainsi que le traitement de la variation linguistique présente un défi considérable en linguistique et en TAL pour les langues peu dotées. L'article explore et compare les différentes approches et méthodologies de lemmatisation et d'annotation linguistique pour un corpus multivariationnel, plus particulièrement pour un corpus combinant une variation diachronique et une variation synchronique [1]. L'annotation d'un corpus peut être réalisée via deux approches :

1. une approche classique dite par règles-dictionnaires, qui fait appel à un ensemble de règles pré-établies associées à un dictionnaire des formes annotées. Cette approche nécessite un temps de développement très important et les règles créées sont propres à un état de langue, sur lequel elle est très performante (Dereza, 2018) ;
2. une approche avec des réseaux de neurones, entraînés sur des corpus déjà annotés et qui peuvent réaliser des prédictions sur un nouveau corpus. Les RNN permettent en particulier de réaliser des annotations sur des tokens inconnus, qui sont prédominants dans le cas de corpus très différents, et de proposer des annotations contextuelles (Dereza, 2018). Il s'agit d'une approche gourmande en ressources et en données, mais qui, dans le cas de langues peu dotées, peut constituer une alternative intéressante et efficace à l'approche par règles-dictionnaires.

Jusqu'à présent, l'annotation linguistique (lemmatisation, POS-tagging, annotation morphologique et lexicale) des différentes variétés de l'arménien a principalement été réalisée avec des approches par règles-dictionnaires. Ces approches ont prouvé leur efficacité lorsque le système descriptif est suffisamment complet (dictionnaire), précis (règles) et le corpus traité homogène du point de vue de la variation linguistique (Khurshudyan et al., 2021) [pour l'arménien oriental], mais elles s'avèrent vite limitées et perturbées par différents facteurs identifiés (Vidal-Gorène and Kindt, 2020) [pour l'arménien classique]. Bien que souvent pertinentes, les approches par règles-dictionnaires souffrent de l'important temps de développement, et donc de ressources humaines, qu'elles nécessitent et de leur manque de polyvalence, en particulier pour des variations d'une même langue. Constituer des corpus annotés des différentes variations de l'arménien est en enjeu essentiel pour l'étude de cette langue, mais les annoter à la main ou créer un système règles-dictionnaire pour chacun d'entre eux est difficilement envisageable.

La présente recherche vise donc à explorer une alternative reposant sur des RNN pour annoter des variantes de l'arménien avec un même modèle. Cette approche a été favorisée pour sa flexibilité et son déploiement rapide, aussi bien linguistiquement que structurellement, sur des datasets variés. Les RNN permettent en particulier de réaliser des annotations sur des tokens inconnus, qui sont prédominants dans le cas de corpus multi-variantes, et de proposer des annotations contextuelles (Dereza, 2018). Des alternatives reposant sur des RNN ont déjà été testées pour l'arménien oriental (Arakelyan et al., 2018; Yavrumyan, 2019), et pour l'arménien classique (Vidal-Gorène and Kindt, 2020), et obtiennent des résultats très variables selon la spécialisation du registre du texte annoté, la taille de la base de données d'apprentissage et le pourcentage de tokens inconnus. Une tentative d'annotation morphologique de différentes ressources en arménien (issues de corpus morphologiquement annotés ou non) est ainsi

1. Les analyses de l'article reprennent partiellement les résultats d'une plus grande recherche visant à créer un corpus unifié diachronique et variationnel de la langue arménienne et plus particulièrement centrée sur le traitement des dialectes arméniens (Vidal-Gorène et al., 2020).

réalisée dans une démarche comparative avec les approches précédentes, afin d'évaluer le possible réemploi d'un modèle spécifique de lemmatisation et de POS-tagging sur des variantes linguistiques diverses et peu dotées, et les conditions d'une telle réutilisation.

2 La langue arménienne et état de l'art de ses ressources TAL

L'arménien est une langue indo-européenne à alignement nominatif-accusatif. Dans ses variantes modernes, la langue est morphologiquement dotée d'un système nominal essentiellement agglutinant et d'un système verbal plus fusionnel. Syntaxiquement, c'est une langue à tête finale, dont l'ordre des mots est flexible au niveau de la proposition (SVO / SOV).

La langue arménienne connaît d'importantes variations, que l'on peut schématiquement représenter par un axe horizontal diachronique allant de l'arménien classique (V^{e}-XII^{e} siècles) et l'arménien moyen (XI^{e}-XVII^{e} siècles) à l'arménien moderne (XVII^{e} siècle – aujourd'hui), et par un cercle décrivant le continuum synchronique arménien composé des deux standards de l'arménien moderne (oriental et occidental, standardisés au 19e siècle), de nombreux dialectes et variétés vernaculaires [pour plus d'informations sur la variation linguistique de l'arménien et son contexte géographique voir (Donabédian, 2018; Donabédian and Sitaridou, 2021)].

L'intercompréhension entre les variétés concernées peut varier d'une intelligibilité partielle à nulle. Ainsi, l'arménien classique est une langue flexionnelle avec une riche morphologie (presque totalement inintelligible pour le locuteur contemporain non-initié), l'arménien moyen est intermédiaire entre l'arménien classique et les variantes modernes, et se caractérise par une prolifération des paradigmes morphologiques, ainsi que l'abondance de lexique emprunté (presque totalement inintelligibles pour le locuteur contemporain), et enfin l'arménien occidental moderne et l'arménien oriental moderne et les trois dialectes choisis ont certains traits convergents, notamment une morphologie plus agglutinante, mais si l'intercompréhension entre les deux standards est relativement importante, a fortiori à l'écrit, les dialectes restent largement incompréhensibles pour les locuteurs des standards.

Bien que doté d'une tradition écrite multiséculaire, l'arménien manque significativement de ressources numériques. Plusieurs projets importants se consacrent au développement de ressources TAL pour des variétés arméniennes spécifiques participant d'une dynamique générale ascendante.

Les ressources existantes sont hétérogènes en termes d'accessibilité, de formatage et d'arrière-plan linguistique, lorsqu'elles existent pour une variante, elles ne la couvrent que très partiellement (à l'exception de l'EANC, qui vise à l'exhaustivité pour l'arménien oriental moderne), la majorité des variantes de l'héritage linguistique arménien sont totalement absentes, et les ressources de base du TAL sont encore plus rare.

Arménien classique : Les ressources les plus importantes de l'arménien classique comprennent la Bibliothèque numérique de littérature arménienne (Digital Library of Armenian Literature - Digilib) [2] avec une base de données textuelle simple, sans annotations linguistiques ; le corpus de la Bible (environ 630 000 tokens avec 60 000 tokens uniques et 12 000 lexèmes) avec annotation morphologique complète et alignement anglais réalisé par la fondation Arak29 [3] ; un corpus linguistique plus modeste (66 812 tokens avec 16 000 tokens uniques) spécialisé dans les textes hellénophiles classiques

2. www.digilib.am
3. www.arak29.am

arméniens (6ème-7ème siècles) établi par le projet GREgORI (UCLouvain) [4], le projet Calfa avec sa plateforme complète de dictionnaire de référence d'arménien classique en ligne (1,3 million de tokens, 190 000 tokens uniques) [5], et plusieurs autres bases de données d'arménien classique (projet TITUS [6], Leiden Armenian Lexical Textbase [7]).

Arménien moyen : Aucun corpus dédié à l'arménien moyen et à l'arménien occidental moderne n'est actuellement disponible.

Arménien moderne occidental : Digilib possède la plus importante base de données en texte brut de fiction et de textes historiques de l'arménien occidental moderne des XIX^{e} et XX^{e} siècles.

Arménien moderne oriental : La plus grande ressource pour l'arménien oriental est le Corpus national de l'arménien oriental (EANC) [8], un corpus exhaustif en libre accès avec environ 110 millions de tokens (du milieu du XIX^{e} siècle à aujourd'hui) et une annotation morphologique complète utilisant une approche fondée sur des règles. Le laboratoire YerevaNN et les UD [9] fournissent un échantillon annoté de l'arménien oriental moderne sous la forme d'une banque complète d'arbres de dépendance (53 000 tokens).

Dialectes arméniens : Le seul corpus dialectal [10] accessible en ligne a été conçu dans le cadre du projet de recherche EANC (environ 40 heures d'enregistrements et 250 000 tokens).

3 Datasets, expériences et méthodologies

Cinq principaux datasets [11] ont été constitués pour nos expérimentations : trois sur des variantes dialectales documentées et transcrites dans le cadre du projet de recherche EANC, et deux correspondant à chacun des standards de l'arménien moderne. Trois datasets mixtes ont été établis pour évaluer les gains potentiels en combinant les ressources. Ces données proviennent de la base de données de EANC.

1. **D-Ab** : Ce dataset est composé de transcriptions orales (15 heures, 16 informants) du dialecte d'Arcvaberd (Shamshadin, région de Tavush). Il comprend 120 258 wordforms (14 405 uniques) annotées manuellement. Il s'agit du dataset dialectal le plus important mais aussi le moins varié, avec seulement 4 120 lemmes uniques. **D-Ab** comporte de très nombreuses formes ambiguës, c'est-à-dire dont l'annotation définitive nécessite un arbitrage en fonction du contexte.
2. **D-Ga** : Ce dataset est composé de transcriptions orales (15 heures, 26 informants) du dialecte de Gusana (Maralik, région de Shirak). Il est composé de 100 352 wordforms (20 647 uniques) annotées manuellement. Avec un volume équivalent à **D-Ab**, il est beaucoup plus varié avec 9 087 lemmes uniques. En conséquence, nous trouvons beaucoup plus de tokens inconnus dans l'ensemble des tests qui lui est associé et **D-Ga** constitue donc un repère intéressant pour

4. www.gregoriproject.com
5. www.calfa.fr
6. www.titus.uni-frankfurt.de/indexe.htm
7. www.sd-editions.com/LALT/home.html
8. www.eanc.net
9. www.universaldependencies.org/treebanks/hy_armtdp/index.html
10. www.web-corpora.net/EANC_dialects/search
11. Pour décrire les résultats, nous définissons les acronymes suivants : **D-** pour dataset, associé à la langue concernée (p. ex. **D-MEA**). De même, les modèles créés sont nommés par : **m-** pour le modèle, associé à la langue concernée (p. ex. **m-MEA**).

l'évaluation des prédictions sur des tokens inconnus.

3. **D-Shn** : Ce dataset, composé de transcriptions orales (15 heures, 18 informants) du dialecte de Shenavan (Aparan, région de Aragatsotn), est le plus réduit des trois datasets. Il est composé de 89 632 wordforms (17 940 uniques) annotées manuellement. Il s'agit proportionnellement du dataset le plus varié, avec 7 568 lemmes uniques, et donc de nombreuses formes ambiguës et inconnues.
4. **D-MEA** : Il s'agit du dataset de référence de cet article, en arménien moderne oriental, sous-ensemble de EANC. Il est composé de 5 111 614 wordforms (201 710 uniques). Les phrases sont issues de sources hétérogènes, de la presse arménienne (2 037 629 wordforms), de fictions (1 453 894 wordforms) et de non-fictions (2 031 055 wordforms). **D-MEA** est représentatif de l'arménien moderne oriental.
5. **D-MWA** : Il s'agit d'un dataset expérimental en arménien moderne occidental, destiné à évaluer la pertinence de la réutilisation et reproductibilité des modèles. Il est composé de 3 531 wordforms (1 788 uniques).

Trois datasets mixtes ont été créés : **D-Ab+MEA**, composé de **D-Ab** et augmenté d'un tiers de son volume avec des données variées de **D-MEA**, ainsi que **D-Ga+MEA** et **D-Shn+MEA** augmentés selon le même procédé.

Enfin, trois autres datasets externes ont été considérés pour la création de modèles : les Universal Dependencies (**D-UD**) pour l'arménien oriental, et les données de GREgORI (**D-CA1**) et d'Arak29 (**D-CA2**) pour l'arménien classique.

Les annotations pour **D-Ab**, **D-GA** et **D-Shn** sont réalisées hors contexte (application d'une simple correspondance entre la liste des wordforms et le corpus), contrairement à **D-UD**, **D-CA1** et **D-CA2**. Par ailleurs, le paramètre de la graphie intervient dans la performance des annotations : **D-Ab**, **D-Ga**, **D-Shn**, **D-MEA**, **D-UD** utilisent une orthographe réformée, tandis que **D-MWA**, **D-CA1** et **D-CA2** l'orthographe classique. Des conversions orthographiques ont été réalisées pour permettre l'évaluation : si elles s'avèrent précises au niveau des mots, certaines constructions verbales n'existant pas dans toutes les variations, il en résulte parfois la création de bruit supplémentaire. De plus, **D-Ab**, **D-GA**, et **D-Shn** ont été transcrits en prenant en compte certaines spécificités phonologiques, ce qui ajoute une divergence graphique supplémentaire et affecte l'évaluation des différents modèles.

	D-MEA	**D-Ab**	**D-Ga**	**D-Shn**	**D-MWA**
Mots	5 111 614	120 258	100 352	89 632	3 531
Tokens uniques	201 710	14 405	20 647	17 940	1 788
Lemma uniques	-	4 120	9 087	7 568	1 311
Tokens ambigus	-	18 584	12 883	14 844	250
Tokens inconnus (test dataset)	13 145	364	1 968	1 810	1 080

TABLE 1 – Composition des principaux datasets

Trois types de réseaux ont été entraînés et évalués :

1. modèle RNN univariationnel pour une variété ciblée ;
2. modèle mixte (modèle 2/3 dialecte +1/3 arménien oriental moderne) ;
3. modèle RNN univariationnel pour une variété non-ciblée.

L'architecture RNN utilisée repose sur Pie (Manjavacas et al., 2019), qui offre une architecture très modulaire particulièrement adaptée au traitement des langues anciennes et variées, ce qui est majoritairement le cas ici et déjà éprouvée avec succès sur l'arménien classique (Vidal-Gorène and Kindt, 2020), dont nous reprenons l'essentiel du processus. Nous avons limité la capacité d'apprentissage de Pie — qui exploite pleinement le contexte d'une phrase pour améliorer les tâches de lemmatisation et de POS-tagging, en particulier en cas de token ambigu (Eger et al., 2016; Sprugnoli et al., 2020) — en raison de l'ambiguïté non levée des annotations dans **D-Ab**, **D-Ga** et **D-Shn**. Le RNN réalise donc par défaut une prédiction de toutes les analyses possibles pour un mot. La sélection de l'analyse la plus probable intervient seulement dans un second temps avec un modèle de langue.

Concernant le POS-tagging, le décodeur linéaire avait produit des résultats meilleurs sur l'arménien classique, mais nous avons néanmoins de nouveau comparé avec le décodeur CRF fourni par MarMoT et LEMMING (Mueller et al., 2013; Müller et al., 2015) et qui a fait ses preuves sur des datasets équivalents lors de la dernière Evalatin Evaluation Campaign (Sprugnoli et al., 2020; Stoeckel et al., 2020).

Enfin, pour que la comparaison soit complète, nous avons évalué la pertinence de l'architecture utilisée sur une langue standard (arménien oriental moderne - MEA), sur deux datasets (**D-MEA** et **D-UD**). Entraîné avec **D-UD** 2.3, COMBO (Rybak and Wróblewska, 2018) est efficace à 88,05% en lemmatisation et 85,07% en POS-tagging (Arakelyan et al., 2018; Yavrumyan, 2019). Entraînée avec **D-UD** 2.6, l'architecture décrite ici obtient 91,56% en lemmatisation (74,35% pour les tokens ambigus et 61,85% pour les tokens inconnus) et 92,54% en POS-tagging (87,81% ambigus et 83,56% inconnus).

Nous avons obtenu les résultats suivants en lemmatisation (voir annexes figure 2).

Modèles spécialisés : Pour la lemmatisation générale de tous les tokens des dialectes, les résultats varient entre 92,05% et 97,69%. Dans le cas des tokens inconnus, ceux-ci varient grandement, de 46,52% à 66,87%. Le modèle **m-Ab** (entraîné avec **D-Ab** qui contient le plus de tokens) s'avère être le plus performant dans la tâche générale, mais le manque de variété de tokens et de lemmes conduit à de mauvaises prédictions sur des tokens inconnus, là où **m-Ga** et **m-Shn** s'avèrent plus robustes. La matrice de confusion montre que **m-Ab** échoue majoritairement sur les formes verbales qui ne présentent pas de particularité phonétique (dans la transcription). **m-Ga** et **m-Shn** génèrent quant à eux beaucoup plus de formes fautives pour un même token, en plus d'être pénalisés par la grande variété de prononciations reproduites dans les corpus. **m-MEA** est quant à lui très robuste mais souffre de l'ambiguïté de ses données. Conséquence : le modèle s'avère plus performant quand on lui demande de générer toutes les analyses possibles plutôt que d'une seule (contrairement à **m-UD**). Il est néanmoins efficace à 94,34% sur **D-UD**.

Modèles mixtes : L'ajout de données de **D-MEA** à **D-Ab**, **D-Ga** et **D-Shn** en apprentissage entraîne un gain certain pour le dialecte d'Arcvaberd (+ 0,64% en accuracy, mais surtout +4,4% en précision et recall), y compris pour la prédiction sur des tokens inconnus qui passe de 46,52% à 51,10%. En revanche, cela pénalise les dialectes de Gusana et Shenavan (voir infra Modèles non-spécialisés pour une piste d'explication).

Modèles non-spécialisés : Un modèle strictement entraîné sur l'arménien oriental (**m-MEA** et **m-UD**) ne permet pour l'instant pas une lemmatisation des dialectes. Il en est de même pour des modèles entraînés sur l'arménien classique (**m-CA1** et **m-CA2**). Néanmoins, ces résultats sont à nuancer, puisque **D-Ab**, **D-Ga** et **D-Shn** sont transcrits très différemment et en orthographe réformée.

De nouvelles expérimentations doivent être réalisées après une uniformisation des annotations. Linguistiquement proches, les dialectes Arcvaberd et Gusana n'atteignent pas les 50% de bonne lemmatisation (respectivement 49,47% et 46,38%), ce qui reste néanmoins meilleur que l'annotation de **D-Ab** par **m-Shn** (42,90%). **m-Shn** annote correctement **D-Ga** à 58,45%, tandis que **m-Ga** annote **D-Shn** à 52,32%. On remarque un gain lors de l'utilisation de modèles mixtes.

Compte tenu des limites exposées dans la lemmatisation, les résultats en POS-tagging sont beaucoup plus réguliers (voir annexes figure 3). Il n'a pas été possible de réaliser toutes les évaluations de POS-tagging en raison de la trop grande variation d'annotation des corpus (étiquettes utilisées et caractérisation). Ce dernier point constitue un enjeu important pour des expérimentations ultérieures.

Modèles spécialisés : Nous obtenons de très bons modèles de POS-tagging (> 95%) pour chacun des dialectes, y compris dans le cas de tokens inconnus. **m-Ab** reste peu robuste face à la diversité. Plus de deux tiers des erreurs sont localisés sur une confusion entre nom et adjectif, qui peut s'expliquer par l'absence de contexte.

Modèles mixtes : En POS-tagging, l'ajout de MEA aux dialectes apporte un vrai gain dans l'annotation, en particulier pour les tokens inconnus.

Modèles non spécialisés : La réutilisation des modèles d'un dialecte à un autre apparaît clairement possible ici, en particulier pour les dialectes Gusana et Shenavan, qui n'appartient pourtant pas à la même branche linguistique, avec une couverture de 82,22% de **D-Ga** par **m-Shn+MEA**. Par ailleurs, même seulement efficace à 66,27% (**m-Ga** sur **D-Ab**), cela peut apporter une bonne base pour l'annotation plus rapide de corpus dialectaux.

Le modèle **m-MEA** propose une lemmatisation correcte à 88,79% et un POS-tagging correct à 87,33% d'un corpus en arménien occidental moderne. Le parser de EANC obtient respectivement 74,09% et 68,57% sur ce même corpus.

4 Conclusion

Nous avons réalisé différentes expérimentations qui illustrent pour la première fois l'annotation automatique de variantes dialectales de l'arménien, et la réutilisation possible de modèles non-spécialisés pour la constitution rapide de corpora.

La précision moyenne d'annotation des RNN est de 94% en lemmatisation et 97,02% en POS-tagging. Ils ont notamment montré une grande polyvalence pour traiter différentes variétés linguistiques autres que leur base d'apprentissage, au contraire des systèmes traditionnels. Les tests réalisés mettent notamment en évidence une reproductibilité des modèles de l'ordre de 74%. Mais il a toutefois fallu les entraîner avec une base d'apprentissage préalablement annotée contenant plus de 5 millions de tokens. Entraînés sur des corpus plus modestes (60 000 tokens), les RNN proposent des résultats parfois en-deça des systèmes règles-dictionnaires (précision moyenne de 92%), bien que restant meilleurs sur les tokens inconnus (Vidal-Gorène and Kindt, 2020). En particulier, la comparaison d'un modèle RNN entraîné sur de l'arménien oriental et d'un modèle règles-dictionnaire en arménien oriental appliqués à de l'arménien occidental moderne montre un gain significatif de 19% en annotation. Le modèle RNN couvre en effet 88,79% d'un petit ensemble de données hétérogènes en arménien occidental, modèle qui pourrait donc être un point de référence pour l'annotation massive de corpus dans ce standard linguistique.

Nous soutenons ainsi qu'un modèle reposant sur des RNN peut être une alternative probante à une approche par règles, compte tenu de facteurs comme le temps de traitement, la réutilisation d'un même modèle sur différentes variétés d'une langue donnée, et les gains en annotation morphologique pour des langues peu dotées. Cela est d'autant plus pertinent dans le cadre d'une langue peu dotée, car il permet de construire rapidement une base d'apprentissage suffisante pour établir un RNN polyvalent. Les différentes études conduites, en complément de la mise à disposition des architectures neuronales ou hybrides construites, servent de point de repère pour l'annotation de langues peu dotées à graphies non latines. Une nouvelle approche, hybride, devrait pouvoir concilier rapidité d'application et couverture maximale de l'annotation. La présente recherche montre une possibilité de réutiliser des modèles pour couvrir d'autres variétés linguistiques, même partiellement.

Des recherches futures permettront d'étendre le corpus multivariationnel afin d'inclure autant de variantes linguistiques arméniennes que possible (des corpus visant à l'exhaustivité pour l'arménien classique, l'arménien moyen et l'arménien occidental moderne, ainsi que des dialectes arméniens et des variantes vernaculaires) et de tester et évaluer les modèles existants sur les nouvelles données.

Jusqu'à présent, la distance variationnelle entre variantes linguistiques a été calculée sur la base d'un faisceau de traits linguistiques. La question de la distance linguistique entre deux variétés est particulièrement pertinente dans la classification généalogique des langues du monde et dans le cadre de la classification des dialectes. Dans la tradition dialectologique arménienne, on a eu recours pour la classification des dialectes à divers critères linguistiques et extra-linguistiques (géographique, morphologique, phonologique, multiparamétrique, etc.). Appliquer le modèle d'annotation morphologique RNN d'une variété particulière sur d'autres variétés typologiquement proches (dychroniquement ou synchroniquement) et évaluer la distance variationnelle entre deux variétés sur la base de différents paramètres formels permettrait de formuler des hypothèses nouvelles et de revoir les classifications existantes.

Les variétés linguistiques ont généralement une vitalité fragile en l'absence de prestige social et / ou de renouvellement des locuteurs natifs. C'est en particulier le cas pour les dialectes arméniens qui souffrent d'un déficit de reconnaissance par rapport à la langue standard avec laquelle ils coexistent. Par conséquent, la documentation et l'annotation de ces variétés linguistiques sont d'un enjeu majeur en TAL mais aussi et surtout dans des perspectives linguistiques, anthropologiques et sociales.

Références

Arakelyan, G., Hambardzumyan, K., and Khachatrian, H. (2018). Towards JointUD : Part-of-speech Tagging and Lemmatization using Recurrent Neural Networks. In *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 180–186, Brussels, Belgium. Association for Computational Linguistics.

Dereza, O. (2018). *Lemmatization for Ancient Languages : Rules or Neural Networks ?*, pages 35–47. Springer International Publishing, Cham.

Donabédian, A. (2018). Middle East and Beyond - Western Armenian at the crossroads : A sociolinguistic and typological sketch. In Bulut, C., editor, *Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery*, volume 111 of *Turcologica*, pages 89–148. Harrazowitz Verlag, Wiesbaden, Allemagne.

Donabédian, A. and Sitaridou, I. (2021). Anatolia. In Adamou, E. and Matras, Y., editors, *The Routledge Handbook of Language Contact*, pages 404–433. Routledge, London, England.

Eger, S., Gleim, R., and Mehler, A. (2016). Lemmatization and Morphological Tagging in German and Latin : A comparison and a survey of the state-of-the-art. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1507–1513, Portorož, Slovenia. European Language Resources Association (ELRA).

Khurshudyan, V., Arkhangelskiy, T., Daniel, M., Levonian, D., Plungian, V., Polyakov, A., and Rubakov, S. (2021). Introduction to Eastern Armenian National Corpus : www.eanc.net. *Études arméniennes contemporaines*. submitted.

Manjavacas, E., Ákos, K., and Mike, K. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Rybak, P. and Wróblewska, A. (2018). Semi-Supervised Neural System for Tagging, Parsing and Lematization. In *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium. Association for Computational Linguistics.

Sprugnoli, R., Passarotti, M., Cecchini, F. M., and Pellegrini, M. (2020). Overview of the EvaLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Stoeckel, M., Henlein, A., Hemati, W., and Mehler, A. (2020). Voting for POS tagging of Latin texts : Using the flair of FLAIR to better Ensemble Classifiers by Example of Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135, Marseille, France. European Language Resources Association (ELRA).

Vidal-Gorène, C., Khurshudyan, V., and Donabédian, A. (2020). Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 90–101, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Vidal-Gorène, C. and Kindt, B. (2020). Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old georgian, and Syriac. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 22–27, Marseille, France. European Language Resources Association (ELRA).

Yavrumyan, M. (2019). Tokenization and Word Segmentation in the UD ARMENIAN-ArmTDP Treebank. *Banber Erewani hamalsarani*, pages 52–65.

5 Annexes

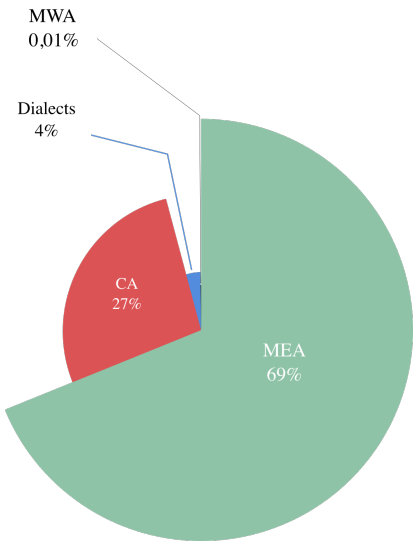

FIGURE 1 – Proportion par langue des ressources annotées et en open access, utilisées dans le cadre de cette étude.

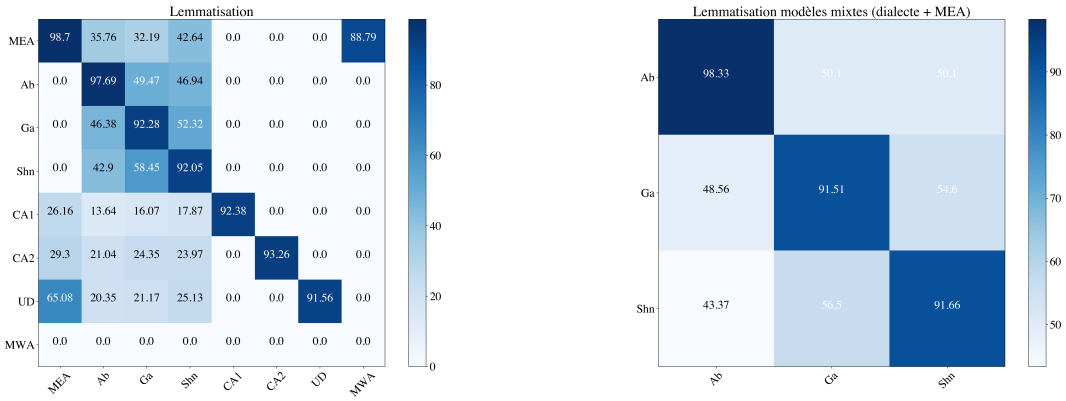

FIGURE 2 – Évaluation des modèles simples et mixtes de lemmatisation sur tous les tokens (accuracy) répartition des résultats. La valeur 0 indique que le modèle n'a pu être appliqué sur le dataset concerné, en raison des limites évoquées dans l'article.

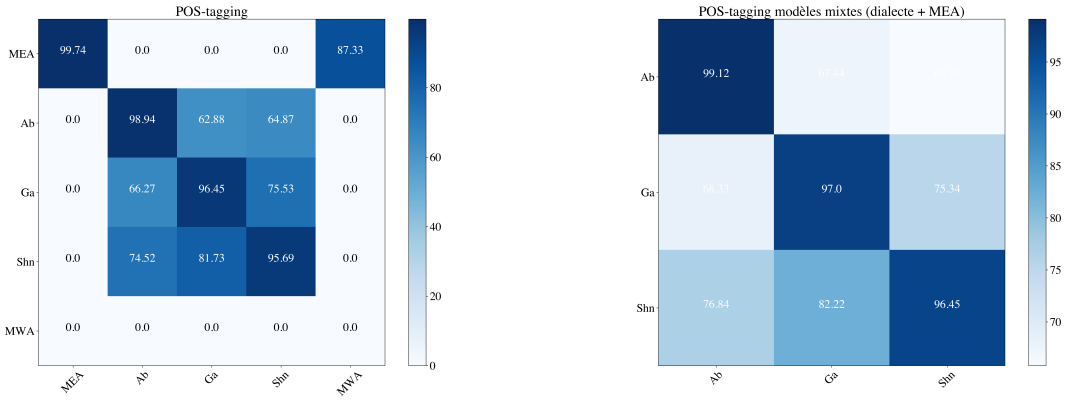

FIGURE 3 – Évaluation des modèles simples et mixtes de POS-tagging sur tous les tokens (accuracy) et répartition des résultats. La valeur 0 indique que le modèle n'a pu être appliqué sur le dataset concerné, en raison des limites évoquées dans l'article.

Ouvrir aux linguistes « de terrain » un accès à la transcription automatique

Guillaume Wisniewski[1], Alexis Michaud[2], Benjamin Galliot[2], Laurent Besacier[3], Séverine Guillaume[2], Katya Aplonova[4] et Guillaume Jacques[5]
(1) Laboratoire de linguistique formelle (LLF), CNRS-Université de Paris, France
(2) Langues et civilisations à tradition orale (LACITO), CNRS-Sorbonne Nouvelle, France
(3) Laboratoire d'informatique de Grenoble (LIG), CNRS-Université Grenoble Alpes, France
(4) Langage, langues et cultures d'Afrique (LLACAN), CNRS-INALCO, France
(5) Centre de recherches linguistiques sur l'Asie orientale (CRLAO), CNRS-EHESS, France
guillaume.wisniewski@u-paris.fr, alexis.michaud@cnrs.fr,
b.g01lyon@gmail.com, laurent.besacier@univ-grenoble-alpes.fr,
severine.guillaume@cnrs.fr, {aplooon|rgyalrongskad}@gmail.com

Résumé

Le traitement automatique de la parole commence à réaliser son fort potentiel pour la documentation des langues en danger. Notre objectif est de mettre à la portée des linguistes « de terrain » des outils de transcription automatique à la pointe des avancées technologiques. Une interface graphique conviviale, Elpis, donne désormais accès à Kaldi et ESPnet, deux bibliothèques de pointe pour le traitement automatique de la parole. Une *recette* ESPnet à utiliser dans Elpis donne d'excellents résultats, aussi bien sur deux jeux de données précédemment utilisés pour entraîner des modèles acoustiques (langues na et chatino) qu'avec deux nouveaux jeux de données (japhug et bashkir). L'interface utilisateur d'Elpis a en outre été dotée de traductions. L'installation est facilitée par *conteneurisation* (en utilisant le logiciel libre Docker), et l'entraînement des modèles est accéléré par l'utilisation de processeurs graphiques (à l'aide de la bibliothèque CUDA).

Abstract

User-friendly automatic transcription of low-resource languages

Natural Language Processing now begins to deliver on its promise for language documentation. This paper reports on progress integrating the speech recognition toolkit ESPnet into Elpis, a web front-end originally designed to provide access to the Kaldi automatic speech recognition toolkit. The goal of this work is to make end-to-end speech recognition models available to language workers via a user-friendly graphical interface. Encouraging results are reported on (i) developing an ESPnet recipe for use in Elpis, with preliminary results on data sets previously used for training acoustic models with the Persephone toolkit along with two new data sets that had not previously been used in speech recognition, and (ii) incorporating ESPnet into Elpis along with user interface enhancements and a CUDA-supported dockerfile.

Mots-clés : documentation linguistique, documentation linguistique assistée par ordinateur, reconnaissance automatique de la parole, science ouverte, linguistique de terrain.

Keywords : Language documentation, Computational Language Documentation, Automatic Speech Recognition, Open Science, linguistic fieldwork.

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 82–93,

1 Introduction

La transcription de la parole constitue une dimension importante de la documentation linguistique, en particulier s'agissant de langues et civilisations à tradition orale. Des progrès spectaculaires ont été réalisés en matière de reconnaissance automatique de la parole au cours de la dernière décennie (Hinton et al., 2012 ; Hannun et al., 2014 ; Zeyer et al., 2018 ; Hadian et al., 2018 ; Ravanelli et al., 2019 ; Zhou et al., 2020), y compris de grandes réussites pour des langues peu documentées, pour lesquelles peu de ressources numériques sont disponibles (Besacier et al., 2014 ; Blokland et al., 2015 ; Lim et al., 2018 ; van Esch et al., 2019 ; Hjortnaes et al., 2020). Néanmoins, les technologies de transcription automatique ne sont pas encore exploitées à grande échelle par les linguistes « de terrain » et leurs collaborateurs et collaboratrices. Les logiciels de reconnaissance de la parole sont souvent des prototypes de recherche pour lesquels il n'existe pas d'interface graphique et qui exigent des compétences informatiques peu répandues parmi les utilisateurs potentiels, à commencer par une familiarité avec l'utilisation de la ligne de commande.

L'enjeu pour la documentation linguistique est de taille. En effet, la mise en œuvre, dans le cadre d'archives ouvertes telles que la collection Pangloss (Michailovsky et al., 2014), d'outils de pointe en traitement automatique des langues naturelles apporte une forte impulsion à l'enrichissement des ressources hébergées par ces archives. Le dépôt en archive ouverte, auquel les chercheurs ont accès dans le cadre du dispositif mis en place sous l'égide d'Huma-Num (Jacobson et al., 2015), présente des avantages décisifs en termes de pérennité, mais nombre de chercheurs sont (de façon bien compréhensible) plus sensibles aux enjeux de recherche, à court et moyen terme, qu'aux questions qui concernent la postérité lointaine. L'existence de traitements automatisés pour les corpus déposés dans les archives sonores encouragerait chez les linguistes de terrain un changement d'attitude, dans le sens d'un passage à l'*archivage progressif*. En effet, des archives bien outillées en logiciels de TAL ne prêteraient plus le flanc au soupçon de devenir des « cimetières de données » (*data graveyards :* Gippert et al., 2006, 4, 12-13). Elles deviendraient plutôt des « cliniques de données » (*data clinics*)[1], qui offrent aux déposants un soutien technologique dans l'enrichissement de leurs données. (Un exemple en est fourni par les réalisations présentées aux présentes Journées scientifiques par Cécile Macaire, voir sa communication *Alignement temporel entre transcriptions et audio de données de langue japhug*.) Cela encouragerait les dépôts, parce qu'il y aurait à la clef un fort potentiel d'amélioration des transcriptions et annotations (gloses, traductions). De la sorte, nombre de déposants potentiels seraient portés à franchir le pas et engager une démarche d'archivage, plutôt que remettre à plus tard. Le fait de reporter l'archivage, dans l'attente d'un degré de perfection que les corpus n'atteignent souvent jamais, aboutit en effet à une très forte déperdition de données de langues rares.

Elpis[2] est précisément un outil conçu pour permettre aux linguistes et à leurs collaboratrices et collaborateurs d'avoir accès à un outil de reconnaissance automatique de la parole. Il permet d'entraîner son propre modèle acoustique pour la reconnaissance vocale et de transcrire automatiquement des fichiers audio, au moyen d'une interface graphique (Foley et al., 2018, 2019 ; Adams et al., 2021). Le premier moteur de reconnaissance automatique de la parole « historique » auquel donnait accès Elpis était Kaldi[3] (Povey et al., 2011). Kaldi est une bibliothèque libre, très utilisée dans la communauté de traitement de la parole, qui repose sur une architecture qui hybride modèles de Markov cachés et réseaux de neurones profonds. Elle a permis d'obtenir des résultats à la pointe du progrès dans

1. L'expression est d'un relecteur anonyme pour le colloque Comput-EL4 : *Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

2. `https://github.com/CoEDL/elpis`

3. `https://github.com/kaldi-asr/kaldi`

nombre de tâches de reconnaissance de la parole. Dans ce travail, nous décrivons le processus par lequel une autre bibliothèque de reconnaissance de la parole, plus récente, ESPnet[4] (Watanabe et al., 2018), a été ajoutée à Elpis, aux côtés de Kaldi. La raison pour intégrer ESPnet à Elpis est qu'il s'agit d'un outil largement utilisé, qui rassemble une communauté grandissante et fait l'objet d'un développement logiciel soutenu. Mais surtout, ESPnet implémente une architecture de reconnaissance de la parole plus récente reposant uniquement sur des réseaux de neurones : c'est une approche pouvant être qualifiée de bout en bout (*end-to-end*). Le choix entre plusieurs moteurs *(back-ends)* peut permettre d'obtenir de meilleures performances selon la nature du jeu de données. En effet, nombre de locuteurs, nature des documents, rapport signal/bruit, débit de parole... diffèrent considérablement d'un corpus de langue rare à l'autre, de sorte qu'une grande flexibilité des outils est souhaitable.

Nous commencerons par décrire les changements apportés au logiciel Elpis pour y intégrer le moteur ESPnet et le développement d'une recette ESPnet adaptée aux corpus de langues rares. Une recette est un ensemble de scripts et de fichiers de configuration qui permet de faciliter les différentes étapes de la reconnaissance automatique (préparation des données, extraction de traits, entraînement du modèle, etc.), regroupés dans un script enrobeur (*wrapper*) dans lequel il est aisé de spécifier l'architecture du modèle et ses différents hyperparamètres. Nous décrirons ensuite la mise en œuvre de cette recette pour quatre langues. Enfin, nous discuterons des perspectives pour l'avenir de ce projet.

2 Aperçu des outils et des résultats

2.1 Développements réalisés

L'interface d'Elpis permet de charger des fichiers : voir la Fig. 1. Cette interface a été dotée d'une version française ; il est prévu d'ajouter d'autres langues, en collaboration avec des linguistes de terrain qui ont collecté et transcrit des corpus. Au fil de l'entraînement d'un modèle acoustique (Fig. 2), puis de la transcription de nouveaux fichiers audio (Fig. 3), l'interface montre le journal *(log)*. Même si le détail des messages n'est pas intelligible aux utilisateurs, ils peuvent du moins s'assurer que le processus (qui peut durer jusqu'à plusieurs jours) suit son cours et ne s'est pas bloqué. Si l'outil est déployé par une équipe qui possède des compétences en apprentissage automatique, le journal peut être utilisé pour affiner les paramètres en vue d'améliorer les résultats.

Elpis affiche en outre la liste des mots qui apparaissent dans les fichiers utilisés comme corpus d'entraînement, et indique le nombre d'occurrences : voir Fig. 4.

Une image Docker a en outre été réalisée afin de faciliter l'utilisation du logiciel. Cette image est un conteneur isolé comprenant le logiciel ainsi que toutes ses dépendances (bibliothèques, autres programmes, données, etc.), elle peut être téléchargée ou construite localement (selon une suite d'étapes automatisées). Cette image permet notamment de tirer parti des processeurs graphiques (GPU) afin d'accélérer l'entraînement des modèles grâce à l'utilisation d'une bibliothèque de calcul adaptée (CUDA).

La recette ESPnet, développée spécifiquement pour des corpus de petite taille, est librement disponible en ligne, de même que l'ensemble des outils et données[5]. Les données na et japhug, hébergées

4. https://github.com/espnet/espnet

5. https://github.com/persephone-tools/espnet/commit/1c529eab738cc8e68617aebbae520f7c9c919081

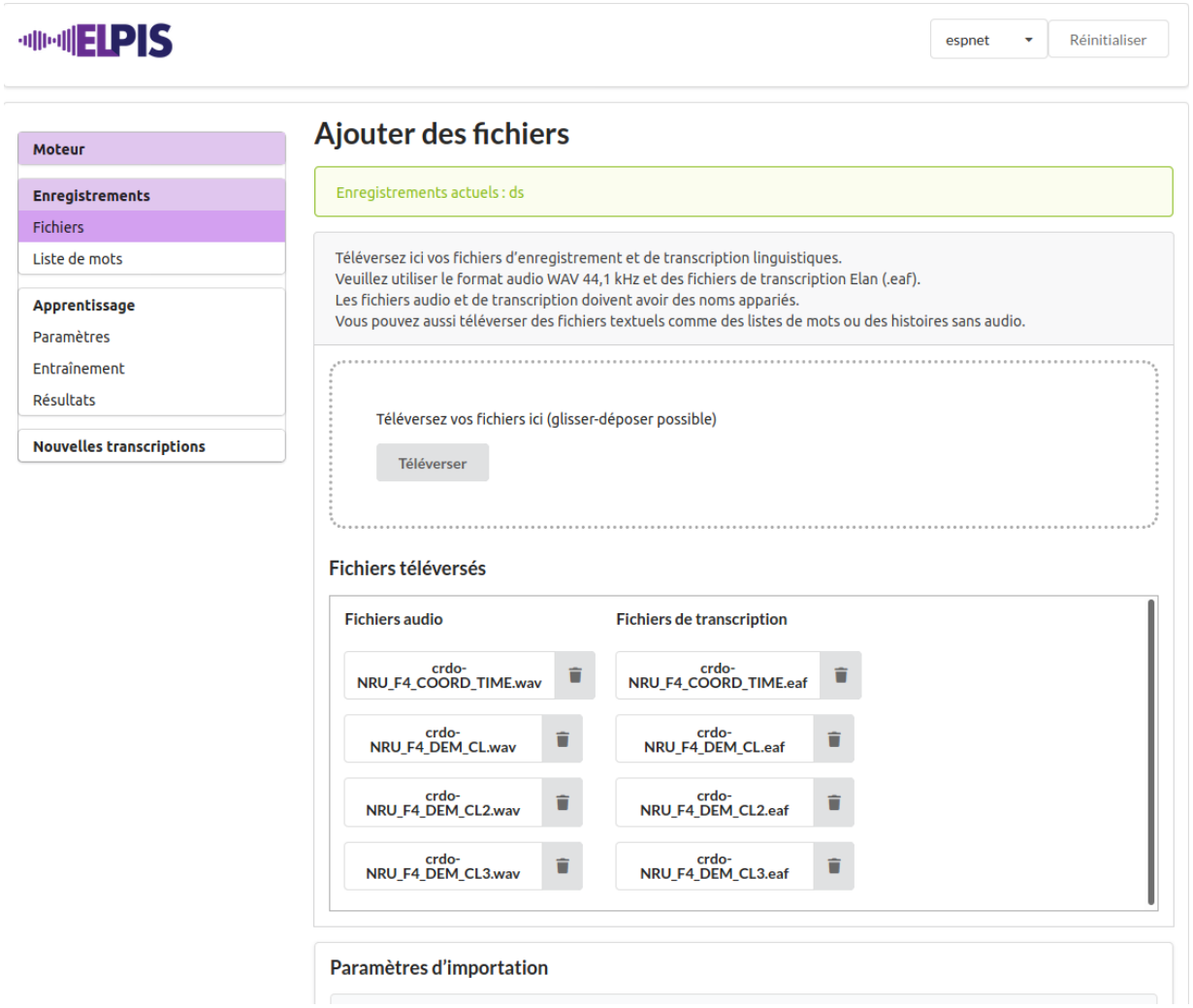

FIG. 1 : Interface d'Elpis. On notera, en haut à droite, le choix de moteur : Kaldi ou ESPnet.

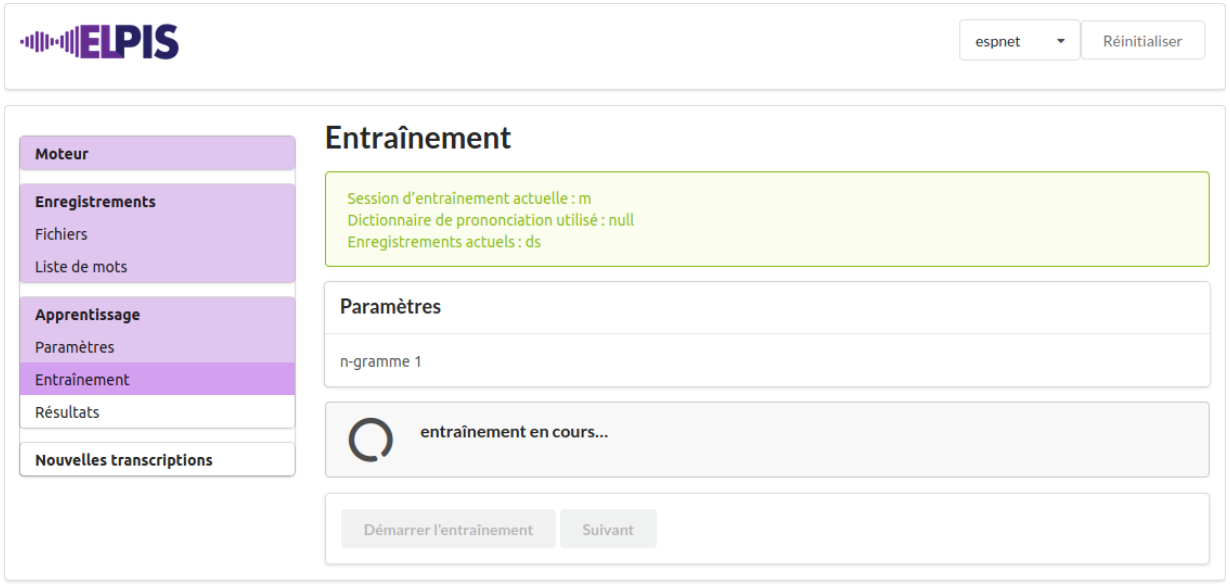

FIG. 2 : Affichage pendant que l'entraînement est en cours.

dans la collection Pangloss, sont converties au format ELAN (seul format d'entrée considéré par Elpis pour le moment) à l'aide d'un script XSLT, baptisé Pangloss-Elpis[6].

6. `https://gitlab.com/lacito/pangloss-elpis`

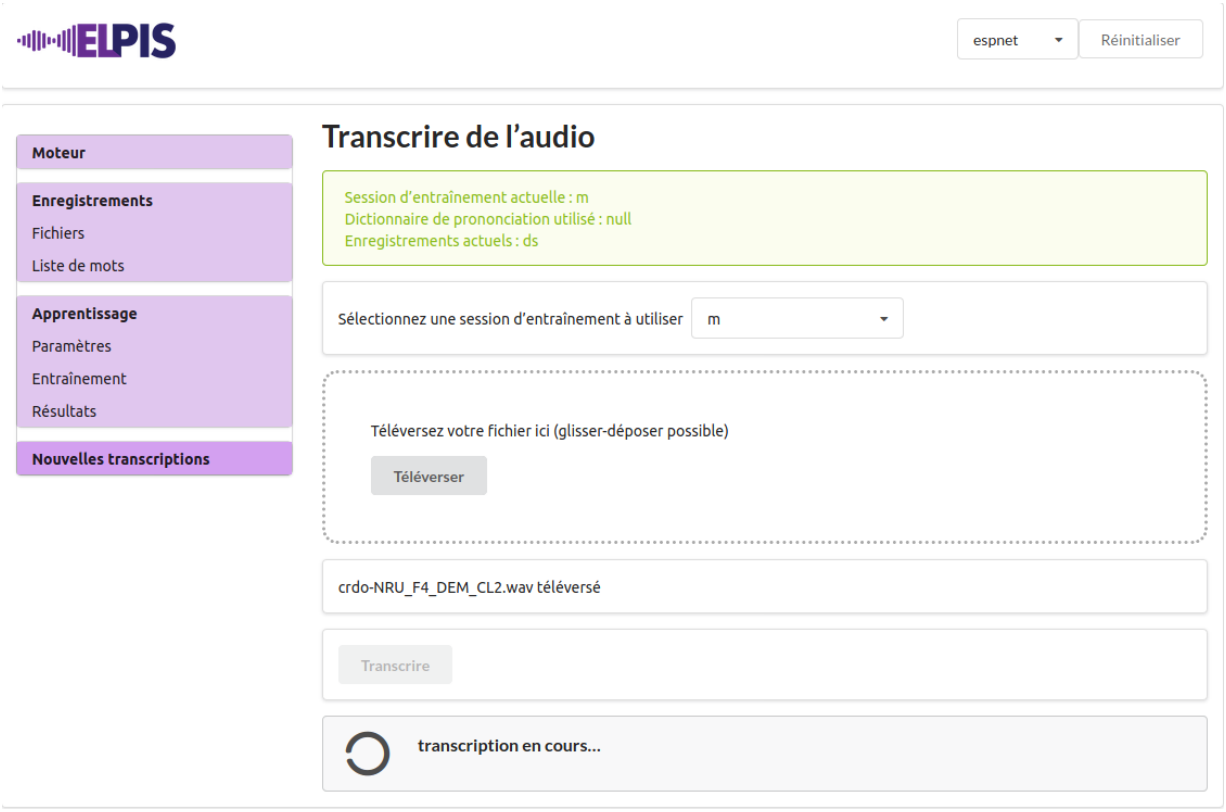

FIG. 3 : Transcription de nouveaux documents audio.

2.2 Résultats

Un bilan des résultats pour quatre langues (cinq jeux de données au total) est présenté dans le tableau 1 ; *taille* désigne la taille du corpus d'entraînement. La figure 5 montre, pour la langue japhug, la baisse du taux d'erreur dans l'identification des phonèmes – ou plus précisément des caractères – selon la taille du corpus d'entraînement, jusqu'à 170 minutes. Des tests sont en cours pour déterminer dans quelle mesure des améliorations sont possibles en utilisant un corpus d'entraînement plus étendu.

Le travail sur le corpus bashkir en est à ses toutes premières étapes et le résultat n'est pas encore probant. Le nombre élevé de locuteurs (36, contre un seul pour les autres corpus : na, chatino et japhug) y est certainement pour beaucoup : le passage du mono-locuteur au multi-locuteurs est une difficulté bien connue pour les systèmes de reconnaissance automatique de la parole, au vu des importantes différences inter-individuelles dans la prononciation. Néanmoins, dans l'ensemble, les résultats obtenus sont encourageants et montrent qu'il est envisageable, dès aujourd'hui, d'utiliser des méthodes de reconnaissance automatique pour faciliter le travail des linguistes de terrain.

Il faut toutefois noter que la qualité des prédictions varie fortement en fonction des langues. Il reste à déterminer dans quelle mesure ces variations tiennent aux propriétés intrinsèques à la langue (taille de l'inventaire phonémique, complexité phonotactique, phénomènes de réduction phonétique liés à la morphosyntaxe, à la structure de l'information...) et dans quelle mesure elles sont liées à des propriétés du corpus (qualité d'enregistrement du signal audio, genres de documents recueillis, nombre de locuteurs). Ce travail prolongera les tests réalisés à ce jour (Adams et al., 2017, 2018 ; Michaud et al., 2018, 2019), qui permettent déjà certaines généralisations (Wisniewski et al., 2020 ; Michaud et al., 2020b,a). De même, des expériences supplémentaires seront nécessaires pour déterminer si une même recette permet d'obtenir des résultats optimaux pour toutes les langues ou s'il sera néces-

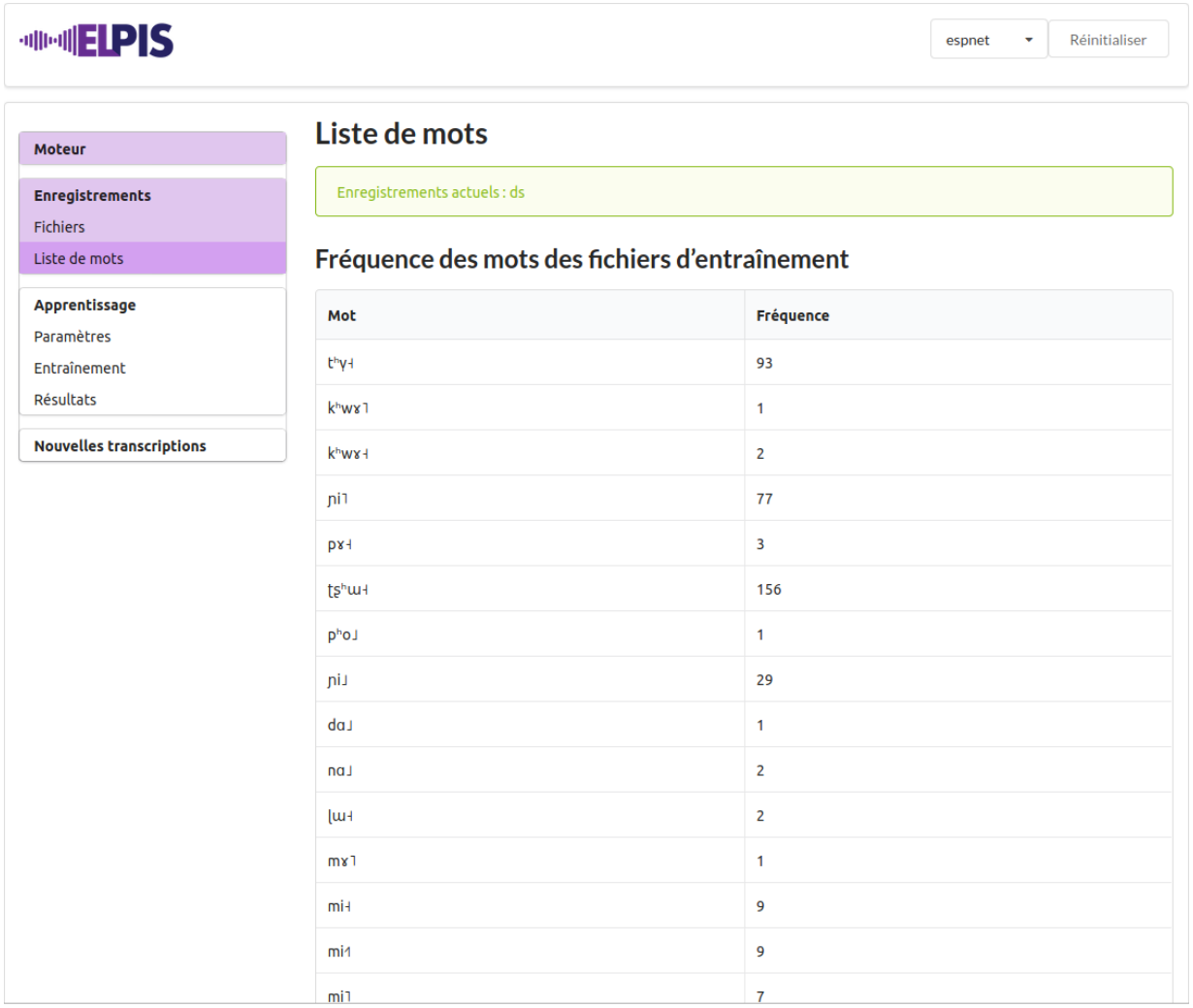

FIG. 4 : Liste des mots du corpus d'entraînement avec mention de leur fréquence d'occurrence.

saire d'adapter l'architecture du réseau de neurones et les hyperparamètres aux spécificités de chaque langue.

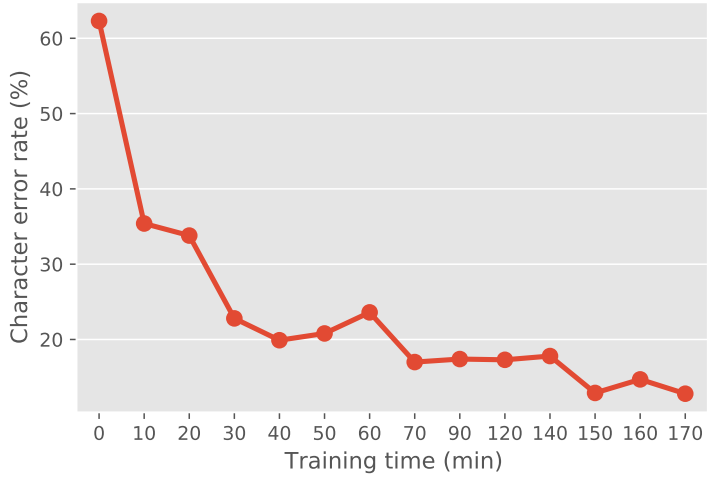

FIG. 5 : Taux d'erreur des graphèmes *(Character Error Rate)* pour le corpus japhug en fonction de la taille du corpus d'entraînement. Moteur : ESPnet.

Langue	Nb locuteurs	Type	Taille (mn)	CER (%)
Na	1	*Récits spontanés*	273	14.5
Na	1	*Expressions élicitées*	188	4.7
Chatino	1	*Parole lue*	81	23.5
Japhug	1	*Récits spontanés*	170	12.8
Bashkir	36	*Récits spontanés*	273	33

TAB. 1 : Résultats de la recette ESPnet sur quatre langues. Les performances sont évaluées par le taux d'erreur des graphèmes *(Character Error Rate)*.

3 Discussion : place d'outils de reconnaissance automatique de la parole dans la documentation linguistique

Un cadre pour une discussion générale au sujet du rôle que peuvent jouer des outils de reconnaissance automatique de la parole dans la documentation linguistique nous est opportunément fourni par l'un des locuteurs invités de la précédente édition des Journées scientifiques du Groupement de recherche LIFT (Orléans, 2019). Dans son exposé, dont une forme développée est sous presse dans la revue *Computational Linguistics* à l'heure où nous écrivons (automne 2020), Steven Bird relève les problèmes que pourrait poser une approche dans laquelle on s'imposerait de généraliser, dans la documentation des langues en voie de disparition, des chaînes de traitement d'outils désormais classiques en reconnaissance de la parole, mais qui seraient mal adaptées aux conditions réelles sur le terrain. La critique de Steven Bird s'adresse notamment à l'emploi d'un modèle acoustique. Passer par l'étape d'une transcription exhaustive en phonèmes (ou autre représentation *segmentale*), au sein de laquelle il faut ensuite identifier les mots, ce serait se créer une difficulté supplémentaire : dans la parole, on ne retrouve pas les phonèmes dans leur réalisation canonique, qui apparaîtraient en une succession sagement linéaire. L'oral est au contraire caractérisé par la variation, les hésitations et reprises, et l'inventivité constante. Steven Bird relève également la diversité des environnements de travail concernés, et des compétences disponibles au sein des groupes qui travaillent à des tâches de documentation linguistique et de sauvegarde des langues. Il préconise une méthode moins linéaire, qu'il dénomme « transcription clairsemée » (« *sparse transcription* ») : annoter ce qui peut l'être aisément, en tirant le meilleur parti des diverses sources d'information disponibles (notamment des traductions), sans s'astreindre d'emblée à une transcription exhaustive et linéaire (Bird, 2020).

Ces réflexions offrent l'occasion de préciser la place qu'occupe un outil tel qu'Elpis dans le paysage des méthodes computationnelles pour la documentation linguistique. Personne ne conteste qu'il serait bien utile de disposer d'une transcription phonémique automatique. En revanche, la question de son intégration dans les chaînes de traitement mérite à l'évidence d'être posée. Steven Bird s'oppose spécifiquement à l'idée selon laquelle la transcription phonémique constituerait une bonne base de départ pour la transcription et l'annotation.

Assurément, la transcription phonémique n'est pas une fin en soi et l'équipe du projet Elpis ne se satisfait pas de parvenir à une transcription en phonèmes : là n'est pas le produit final. L'objectif est évidemment de reconnaître mots et phrases : tel est l'enjeu de la reconnaissance automatique de la parole. Mais il importe de mesurer que transcription phonémique et reconnaissance automatique de la parole ne s'opposent pas : dans les systèmes de reconnaissance purement statistiques de bout en

bout (« end-to-end artificial neural networks »), dont fait partie ESPnet, la distinction n'existe pas au plan technique. C'est là une différence importante, qui distingue ESPnet d'un outil tel que Kaldi. Les méthodes de transcription automatique que nous explorons ne sont pas cantonnées à la reconnaissance d'unités du niveau phonémique. Le fonctionnement du logiciel ESPnet est tout à fait compatible avec un auto-apprentissage *(self-supervised training)* sur des données audio non transcrites, affiné par un ajustement à la langue-cible sur un ensemble de données transcrites (corpus de linguistes tels que ceux utilisés dans notre travail). Or il ne paraît pas absurde, au vu des résultats les plus récents, de penser qu'un outil comme wav2vec 2.0 puisse atteindre des performances remarquables en reconnaissance de la parole (reconnaissance de *mots*) par un processus d'ajustement qui ne repose que sur quelques dizaines de minutes d'enregistrements transcrits de référence, ce qui constitue un ordre de grandeur que savent atteindre les linguistes de terrain. Certes, beaucoup d'expériences ont lieu sur l'anglais, mais il ne nous paraît pas y avoir lieu de douter du potentiel d'extension des mêmes méthodes aux langues rares qui constituent notre objet (voir en particulier Conneau et al., 2020)[7].

En outre, lorsqu'on pèse les avantages respectifs de diverses chaînes de traitement, il est de bonne méthode de tenir compte non seulement des performances actuelles des outils logiciels, mais également de leur potentiel de perfectionnement à court et moyen terme. Avant de conclure qu'il faut se détourner de la chaîne de traitement classique en reconnaissance automatique de la parole (*entraînement sur corpus de référence puis application à la transcription exhaustive de données audio*), il paraît avisé de prendre la mesure des souplesses qu'elle autorise en pratique. Au vu des récents succès de systèmes de reconnaissance de la parole pour des langues en danger (voir en particulier Partanen et al., 2020), il ne paraît pas y avoir lieu de craindre que les chaînes de traitement de la reconnaissance automatique de la parole enferment les linguistes dans une impasse méthodologique.

Clairement, Elpis n'est pas destiné à une utilisation lors des premières étapes de documentation d'une langue. Le choix d'une annotation « clairsemée » (« *sparse transcription* ») paraît particulièrement pertinent dans la situation dans laquelle on se trouve lorsque la langue à décrire n'a pas encore fait l'objet d'une analyse linguistique selon les méthodes de la linguistique de terrain, telles que décrites dans des travaux déjà anciens qui conservent leur actualité (Bouquiaux and Thomas, 1971). Des outils de reconnaissance tels que ceux auxquels Elpis donne accès peuvent en revanche être d'une grande utilité lorsqu'on dispose d'un corpus ciselé par un·e linguiste qui a atteint un certain degré de certitude dans l'analyse de la langue (processus d'analyse qui est toujours en devenir).

En outre, nous faisons l'hypothèse selon laquelle le seuil de taille de corpus à partir duquel une langue peut bénéficier d'outils de traitement automatique de grande qualité va continuer à s'abaisser, de sorte que le cercle des langues pour lesquelles des outils de reconnaissance automatique de la parole puissent être déployé va s'élargir rapidement. Nous pensons que ces progrès pourront être mis à profit quelles que soient les chaînes de traitement adoptées. À mesure du déploiement d'Elpis, nous aurons à cœur de veiller à ce que les utilisateurs ne se trouvent pas enfermés dans le carcan d'une méthode unique, mais que leurs réflexions au fil des expériences contribue au contraire à orienter la manière dont les outils s'adaptent pour répondre aux besoins.

7. L'équipe qui développe l'outil ESPnet étudie actuellement les modalités d'implémentation de ces avancées : voir, par exemple, `https://github.com/espnet/espnet/issues/2609`.

4 Conclusion et perspectives

Les perspectives de déploiement de l'outil paraissent prometteuses. Les projets en cours concernent l'amélioration de la recette et l'amélioration de l'interface. Par ailleurs, des tests sont en cours sur un nouveau jeu de données. En outre, une perspective qui paraît hautement souhaitable serait de proposer Elpis sous forme de service web, sur le modèle de WebMAUS[8]. Les outils développés par l'archive bavaroise de la parole, *Bavarian Speech Archive* (Kisler et al., 2017), ont traité plus de dix millions de fichiers multimédias depuis leur passage en production en 2012, ce qui constitue un modèle de réussite de déploiement à grande échelle d'outils de traitement automatique de la parole.

Remerciements

Le travail décrit ici est réalisé en collaboration par une équipe internationale de chercheurs qui se sont rassemblés au fil du temps autour du projet commun de faciliter les tâches de documentation et description des langues en voie de disparition. Le format du présent document ne permettait pas de faire justice à la liste réelle des contributeurs. Il s'agit au premier chef d'Oliver Adams (qui a identifié ESPnet comme un outil prometteur, et a réalisé son intégration dans Elpis), et de l'équipe du projet Elpis : Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles. Nous remercions également Christopher Cox, qui a réalisé un greffon *(plugin)* de transcription automatique pour le logiciel Elan ; Nick Evans, Nick Thieberger, Steven Morey, qui jouent un rôle important dans la coordination internationale du projet Elpis et sa diffusion auprès des linguistes ; et Hilaria Cruz, Martine Adda, Graham Neubig et Nathan Hill pour leur appui. Faute d'espace pour faire figurer tous les collègues qui mériteraient, à divers titres, d'apparaître parmi les auteurs, nous avons fait le choix de ne retenir comme auteurs du présent exposé en français que les participants au groupement de recherche LIFT, organisateur de l'événement. Choix quelque peu arbitraire, mais qui a à tout le moins recueilli le consentement des autres collaborateurs.

Un grand merci aux collègues et amis consultants des langues concernées par les expériences rapportées ici : pour la langue na, il s'agit en particulier de Mme Latami Dashilame et son fils Latami Dashi ; pour la langue japhug, de Tshendzin (Chen Zhen) ; pour la langue tsuut'ina, remerciements particuliers au Bureau du Commissaire à la langue tsuut'ina.

Nous remercions l'Institut des langues rares (ILARA) de l'École Pratique des Hautes Études, l'Université du Queensland et l'*Australian Research Council Centre of Excellence for the Dynamics of Language* pour le soutien financier apporté au développement d'outils de transcription automatique pour la documentation linguistique. Le présent travail est en outre une contribution au projet Labex « Fondements empiriques de la linguistique » (ANR-10-LABX-0083) ainsi qu'au projet « La documentation computationnelle des langues à l'horizon 2025 » (ANR-19-CE38-0015-04).

Références

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC*

8. https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

2018 (Language Resources and Evaluation Conference), pages 3356–3365, Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648.

Adams, O., Cohn, T., Neubig, G., and Michaud, A. (2017). Phonemic transcription of low-resource tonal languages. In *Proceedings of the 2017 Australasian Language Technology Association Workshop (ALTA 2017)*, pages 53–60, Brisbane, Australia. https://halshs.archives-ouvertes.fr/halshs-01656683.

Adams, O., Galliot, B., Wisniewski, G., Lambourne, N., Foley, B., Sanders-Dwyer, R., Wiles, J., Michaud, A., Guillaume, S., Besacier, L., Cox, C., Aplonova, K., Jacques, G., and Hill, N. (2021). User-friendly automatic transcription of low-resource languages: plugging ESPnet into Elpis. In *Proceedings of ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Hawai'i. https://halshs.archives-ouvertes.fr/halshs-03030529.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Bird, S. (2020). Sparse transcription. *Computational Linguistics*, pages 1–50.

Blokland, R., Fedina, M., Gerstenberger, C., Partanen, N., Rießler, M., and Wilbur, J. (2015). Language documentation meets language technology. In *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages - Septentrio Conference Series*, pages 8–18. http://septentrio.uit.no/index.php/SCS/article/view/3457/3386.

Bouquiaux, L. and Thomas, J. (1971). *Enquête et description des langues à tradition orale. Volume I: l'enquête de terrain et l'analyse grammaticale*. Société d'études linguistiques et anthropologiques de France, Paris, 2nd edition 1976 edition. 3 volumes.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. https://arxiv.org/abs/2006.13979.

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., and Ellison, T. M. (2018). Building speech recognition systems for language documentation: the CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018*, pages 200–204, Gurugram, India. ISCA. https://www.isca-speech.org/archive/SLTU_2018/pdfs/Ben.pdf.

Foley, B., Rakhi, A., Lambourne, N., Buckeridge, N., and Wiles, J. (2019). Elpis, an accessible speech-to-text tool. In *Proceedings of Interspeech 2019*, pages 306–310, Graz. https://www.isca-speech.org/archive/Interspeech_2019/pdfs/8006.pdf.

Gippert, J., Himmelmann, N., and Mosel, U. (2006). Language documentation: What is it and what is it good for. In *Essentials of language documentation*, volume 178, pages 1–30. Walter de Gruyter, Berlin.

Hadian, H., Sameti, H., Povey, D., and Khudanpur, S. (2018). End-to-end speech recognition using lattice-free MMI. In *Interspeech*, pages 12–16. https://danielpovey.com/files/2018_interspeech_end2end.pdf.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*. https://arxiv.org/abs/1412.5567.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Others (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

Hjortnaes, N., Partanen, N., Rießler, M., and Tyers, F. M. (2020). Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37, Wien. Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.iwclul-1.5/.

Jacobson, M., Larrousse, N., and Massol, M. (2015). La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales). In *Archives et données de la recherche (ICA/SUV 2014)*, Paris. http://halshs.archives-ouvertes.fr/halshs-01025106.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347. ISBN: 0885-2308 Publisher: Elsevier.

Lim, K., Partanen, N., and Poibeau, T. (2018). Multilingual dependency parsing for low-resource languages: Case studies on North Saami and Komi-Zyrian. In *Proceedings of LREC (International Conference on Language Resources and Evaluation)*, Miyazaki. https://hal.archives-ouvertes.fr/hal-01856178.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation*, 8:119–135. https://halshs.archives-ouvertes.fr/halshs-01003734.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12:393–429. http://hdl.handle.net/10125/24793.

Michaud, A., Adams, O., Cox, C., and Guillaume, S. (2019). Phonetic lessons from automatic phonemic transcription: preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*, Melbourne. https://halshs.archives-ouvertes.fr/halshs-02059313.

Michaud, A., Adams, O., Cox, C., Guillaume, S., Wisniewski, G., and Galliot, B. (2020a). La transcription du linguiste au miroir de l'intelligence artificielle: réflexions à partir de la transcription phonémique automatique. *Bulletin de la Société de Linguistique de Paris*, 116(1). https://halshs.archives-ouvertes.fr/halshs-02881731/.

Michaud, A., Adams, O., Guillaume, S., and Wisniewski, G. (2020b). Analyse d'erreurs de transcriptions phonémiques automatiques d'une langue « rare » : le na (mosuo). In Benzitoun, C., Braud,

C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole, Traitement Automatique des Langues Naturelles, Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 451–462, Nancy, France. ATALA. https://hal.archives-ouvertes.fr/hal-02798572.

Partanen, N., Hämäläinen, M., and Klooster, T. (2020). Speech recognition for endangered and extinct Samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. https://infoscience.epfl.ch/record/192584/files/Povey_ASRU2011_2011.pdf.

Ravanelli, M., Parcollet, T., and Bengio, Y. (2019). The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6465–6469. IEEE. https://arxiv.org/abs/1811.07453.

van Esch, D., Foley, B., and San, N. (2019). Future directions in technological support for language documentation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, Honolulu, Hawai'i. https://computel-workshop.org/wp-content/uploads/2019/02/CEL3_book_papers_draft.pdf#page=26.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., and Chen, N. (2018). ESPnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*. https://arxiv.org/abs/1804.00015.

Wisniewski, G., Guillaume, S., and Michaud, A. (2020). Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In Beermann, D., Besacier, L., Sakti, S., and Soria, C., editors, *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, pages 306–315, Marseille, France. European Language Resources Association (ELRA). https://halshs.archives-ouvertes.fr/hal-02513914.

Zeyer, A., Irie, K., Schlüter, R., and Ney, H. (2018). Improved training of end-to-end attention models for speech recognition. https://arxiv.org/abs/1805.03294.

Zhou, W., Michel, W., Irie, K., Kitza, M., Schlüter, R., and Ney, H. (2020). The RWTH ASR system for TED-LIUM Release 2: Improving Hybrid HMM with SpecAugment. https://arxiv.org/abs/2004.00960.

RefCo: An initiative to develop a set of quality criteria for fieldwork corpora

Jocelyn Aznar[1], Frank Seifart[1]
(1) ZAS, Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
Schützenstr. 18, 10117 Berlin, Allemagne
aznar@leibniz-zas.de, seifart@leibniz-zas.de

RÉSUMÉ

RefCo: une initiative pour développer un ensemble de critères de qualité pour les corpus de terrain.
RefCo est une initiative issue du projet QUEST dont l'objet est de s'assurer de la qualité des corpus issus de terrains en linguistique. L'objectif est d'établir des critères pour s'assurer que les corpus soient réutilisables, en particulier dans le cadre de recherches comparatives. L'initiative porte à la fois sur les corpus existants, par la mise en place d'un label de qualité, et sur les futurs corpus, par l'élaboration de recommandations et des guides de bonnes pratiques. Le label RefCo est décomposable en deux volets : les métadonnées pour la comparaison linguistique et la documentation du corpus. Enfin, un système de production de citations est également développé afin de faciliter la citation des corpus.

ABSTRACT

RefCo is an initiative of the QUEST project which aims at ensuring the quality of corpora from linguistic fieldwork. The objective is to establish criteria for reusable corpora, especially for comparative linguistic research. The initiative addresses both existing corpora, through the implementation of a quality label, and future corpora, through guidelines for best practices. RefCo's label is currentely composed of two panels: the metadata for cross-linguistic corpora, and the Corpus Documentation. Besides, a system producing citations is also developed in order to facilitate corpus citation.

MOTS-CLÉS : réutilisation, archivage, normalisation, critères de qualité, comparaison linguistique, langues orales, documentation linguistique, corpus de terrain

KEYWORDS: reusability, archive, standards, quality criteria, cross-linguistics, oral languages, linguistic documentation, fieldwork corpora.

1 Introduction

At least since Hale et al.'s (1992) acknowledgment of language endangerment and extinction as a scientific issue, many small languages around the world have been documented in one way or another. These documentations are potentially valuable sources of knowledge about human language and linguistic diversity. But, as the number of language documentations increased, the difficulties for reusing those materials are also becoming more apparent (e.g., Thieberger et al. 2016). In fact, of the vast materials held in repositories such as the Endangered Languages Archive (ELAR), The Language Archive (TLA), or PANGLOSS, only little has been used in linguistics research so far, especially by linguists other than the corpus creators themselves.

To enhance the reusability of such data, the QUEST project (https://cutt.ly/quest-project), funded by the German Ministry of Science from 2019 to 2022, proposes to develop, promote and validate good practices and standards for language documentation corpora through consultation with the research community. The QUEST consortium, as a whole, deals with different types of linguistic data, including multimodal and multilingual data. The project develops generic quality criteria as well as specific recommendations applicable for certain re-use scenarios. These quality criteria will be the basis for a data quality label, which QUEST develops as a mid-term goal, and as guidelines for ongoing and future data collection. Furthermore, QUEST develops a Web interface as a tool to make the QUEST label testing process accessible to a wider audience (Arkhangelskiy, Hedeland, and Riaposov 2020).

The QUEST component RefCo (Reference Corpora) targets research carried out on fieldwork corpora[1], especially comparative research, as one of the QUEST re-use scenarios. By 'fieldwork corpora' we mean sets of monological narratives, both audio and audiovisual, that are typically collected in the context of language documentation projects. These projects could aim at cultural preservation for multiple users, including speech communities, and thus might contain additional data not relevant for RefCo like song recordings. RefCo's objective is to ensure that such fieldwork corpora can be re-used, especially in cross-linguistic studies.

As a first step, QUEST RefCo carries out consultation with stakeholder in the field, including fieldworkers and researchers who carry out comparative research using fieldwork corpora (e.g., Bender et al. 2013; Haig & Schnell 2016; Mettouchi, Frajzyngier & Chanard 2017; Stave et al. 2020). On the basis of that, it develops a set of quality criteria for re-usability which will be implemented as (i) guidelines for fieldworkers, and (ii) as basis for a certificate of cross-linguistic re-usability that can be awarded to corpora based on an evaluation carried out by an expert committee. RefCo thus aims to be relevant to both already existing corpora as well as future documentation projects. Additionally, QUEST RefCo carries out case studies to test the re-usability of data that meets the proposed criteria. As with other QUEST components, RefCo criteria are conceived as standards that are currently accepted by the relevant research community: they can and must be amended and changed when standards of the field evolve. The aim is to define a minimum set of quality criteria, not an extensive one. RefCo criteria focus primarily on the quality of annotations and of metadata, both at the level of individual files and at the level of corpora and to improve the findability of the resources. To achieve this objective, QUEST policy is to promote and facilitate the application of standardized good practices on a dataset, not to enforce the use of as many as possible metadata or annotations. QUEST criteria align standards for fieldwork corpora with internationally accepted standards for research data, including the FAIR principles (Wilkinson et al. 2016), Dublin Core Terms, Schema.org, FOAF, SKOS as well as OLAC (Bird and Simons 2001), a metadata specification dedicated to language resources. To further enhance data reuse, RefCo, as part of the QUEST label, provides bibliographical references which follows the Austin principles (Berez-Kroeker et al. 2018), to ensure that corpus creators are properly credited for their work.

[1] In this article, we interchangeably use the terms "dataset" and "corpus" to refer to the data submitted to RefCo. Within QUEST, a "dataset" is a more abstract concept that allows encompassing the various deposit scenario: a field linguist submitting a corpus on an oral language, a set of legacy recordings, a teacher providing annotated documents. In the case of RefCo, both terms are equivalent as the deposit are mostly fieldwork corpora.

2 RefCo Curation Standards

The RefCo component is a subproject of the QUEST initiative. As such, it inherits the quality criteria defined within the QUEST project for assessing the submitted datasets. In this section, we will focus solely on the quality criteria and process associated with fieldwork corpora and thus RefCo.

2.1 Metadata for Cross-Linguistic Corpora

The Quality Standards for Audiovisual Corpora developed by the QUEST initiative focus primarily on a comprehensive set of metadata for linguistic datasets. Within this set, the RefCo subcomponent defines the 'RefCo module', a set of basic metadata that is currently considered minimal standard by research community of field workers and linguists carrying out comparative research on fieldwork data (e.g. Bowern 2008: 47–62 ; Thieberger and Berez 2011: 105-109; Meakins, Green, and Turpin 2018: 73-78). This includes Glottocode language identification codes for the languages documented as well as the language in which translation and glosses are provided, date and location of recording, speaker age and sex, and that a license for re-use is specified. RefCo explicitly allows for approximate, rather than exact time and age information to bridge archivists' desire for complete and precise metadata with the realities the corpus creators face during fieldwork.

A corpus submitted to RefCo has to be licenced with a Creative Common licence which enables the scientific re-use of the corpus[2]. This is a requirement for corpora that are intended to be used in cross-linguistic studies.

The metadata are to be specified at two levels: the corpus, called dataset in QUEST, which is the entity corresponding to a whole coherent submission by a data creator, typically on one language. A dataset is composed of datapackages, also called sessions or bundles, an abstract entity typically referring in the case of the RefCo to one monological narrative, including media and annotation files. The abstraction of datapackage allows to handle case of a texts distributed over various files, or a file containing various texts.

As many research questions require a minimal amount of data, the number of words that have been transcribed and translated, and of the number of words that are morphologically annotated, have to be specified at the dataset level. We are aware that the number of words is a rough estimate of corpus size, given differences in morphological type and in definitions of words, but we consider that this is a useful approximation.

2.2 Documenting a Corpus and its Conventions

Annotating data with interlinear glossing is a standard practice for linguists working on oral languages that was popularized by Boas (1922). The process is now assisted by using software like ELAN, EXMARaLDA or Anvil and has been the object of some conventions, in particual the Leipzig Glossing Rules. Still, there are considereable variations in the interlinear glossing practices of linguists, as apparent in corpora we are currently processing for RefCo. Therefore, RefCo requires that the information relevant for re-use but that cannot be gleaned which explicitely from the corpus

[2] It includes thus the six derivated licenses associated with Creative Commons, including the non commercial and non derivative onr which will still allow comparative researches to be made. Other Open or Free licenses will be evaluated on the demand of a corpus submitter.

itself or the metadata to be provided in a set of separate documents that we call "Corpus Documentation"[3]. The redaction of this Corpus Documentation involves both the corpus creator and the RefCo component into checking the coherency of the annotation. Crucially, RefCo does not requires the use of one of the other glossing convention, but explicit description of the specific glossing conventions used, including, e.g., abbreviations used in glossing and punctuation.

The Corpus Documentation must start with an ***Overview*** section wich provides information about the types of files present in the corpus and their format. It also asks about the number of items present (texts, transcription units, tokens, glosses and POS tags). The following section, ***Annotation levels***, specifies whether all the files in the corpus respect or not the same conventions, the authors has to provide information concerning the tiers and their names, the way they were segmented and in which language they are written. QUEST requires identification of the (anonymous) speaker IDs. Regarding the ***Transcription***, the author of the corpus has to provide first a table associating each grapheme with their phonological value. If employed, it is important as well to specify all the particular strategies used to transcribe noise, cough, laugh or other paralinguistical speech. The language used for the ***Translation*** has to be specified as well as whether the translation was provided by a native speaker of the destination language, a more vehicular language. If relevant, a glossary explaining the untranslated words in the corpus should be provided. The ***Morphemes*** section is dedicated to the description of choices made to handle morpheme boundaries and non-linear morphology, that is if Leipzip Glossing Rules were applied and which morpheme separators were used. In the ***Glosses***, if the punctuation was used in that tier, its meaning has to be described. The grammatical abbreviations are explicited here as well. If the corpus contains a layer dedicated to the annotation of Part-of-Speech **POS**, their meaning has to be provided here. Finally, if the corpus providers find it relevent, in the **Other** section, they can provide additional information and comments.

3 Benefits of using RefCo

Submitting a corpus to RefCo is, for the corpus creator, a consequent effort. Still, we believe that the RefCo label provides substantial benefits to the corpus submitter and the other different stakeholder in the field, which will incentivize the submission of data sets.

3.1 For the Corpus Submitter

One of the first incentise comes from the submission process itself. As the QUEST RefCo supports field workers by guiding them through consistenty and completeness checks of their metadata and annotations, the quality of their work is be positively affected.

Making the corpus available to the public through the QUEST labelling process adds to the accountability of the fieldworker's research by allowing for replication of results (Riesberg 1998). Publication of datasets on which results are based is increasingly viewed as important and enforced by publication outlets in linguistics, following practices in other sciences. It facilitates fair recognition of the efforts invested by the fieldworker by facilitating proper use and citation of the data. It also facilitates future re-use of the data by the fieldworker herself, as it implies fully and consistently processing and properly archiving data.

[3] The Corpus Documentation and its template, which explains how to write the document, were designed by Kilu von Prince.

3.2 Funding organizations

From a funding organization's perspective, RefCo provides a yardstick for the success of a project that involves data collection. First, there is a growing consensus that the public has certain access rights regarding the re-use of public funded researches and their results (Wilkinson et al. 2016). Existing private funding schemes for fieldwork projects, like ELDP, have implemented schemes to enforce best practices for ELDP-funded fieldworkers and require their grantees to implement those, or the funding would not be resumed (Holton & Seyfeddinipur 2018).

3.3 To the Linguistic Community

As it has been stated many times, every language offers a unique window into the puzzle of the human intellectual capacity. In this sense, making fieldwork data on under-described languages accessible contributes to basing linguistic theorizing on a greater sample than the relatively few, well-described languages that most current theories rely on (Hale et al. 1992; Anand, Chung & Wagers 2015; Norcliffe, Harris & Jaeger 2015). One of the primary concern of QUEST is to render linguistic corpora available to other linguists as well as the scientific community in general. The RefCo QUEST criteria focus on information are of particular relevance to re-use for typology. Howeverr, consistently described, coherent, and accessible datasets resulting form RefCo labelling process will render these accessible also for anthropologists and social scientist in general.

3.4 The Speech Communty

Finally, for a speech community which have been subject of a documentation project labelled by RefCo, the process ensures its sustainability by making them findable and accessible. It raises the visibility of the language contributes at the conservation of the cultural heritage, in particular in case of languages that are endangered of becoming extinct, as a whole or with respect to certain traditional genres (Mosel 2006).

4 Conclusion

The QUEST project is an implementation of the current standards and good practices coming from the archiving and linguistics communities. As multiplying metadata and recommendations regarding language documentation corpora adds to the difficulties encountered by corpus creator during the production of their data, the perspective we adopted is to facilitate their application by data creator. To do so, we provide guidelines for describing the corpus The quality criteria and metedata we have seen in this article associated with RefCo, a subproject of QUEST dedicated to the problematic of corpora for cross-linguistics. The position adopted by RefCo is to ease the production of a corpus first by limiting to the minimum the quality criteria and metadata the dataset creators have to provide, second by accompanying them during that application by providing support and guidelines.

References

Anand, Pranav, Sandra Chung & Matthew Wagers. 2015. *Widening the Net: Challenges for Gathering Linguistic Data in the Digital Age*. Paper submitted to the National Science Foundation as part of its SBE 2020 planning activity. https://www.nsf.gov/sbe/sbe_2020/2020_pdfs/Wagers_Matthew_121.pdf.

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey & Fei Xia. 2013. Towards Creating Precision Grammars from Interlinear Glossed Text: Inferring Large-Scale Typological Properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 74–83. Sofia, Bulgaria: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-2710.

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey & Fei Xia. 2013. Towards Creating Precision Grammars from Interlinear Glossed Text: Inferring Large-Scale Typological Properties. *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 74–83. Sofia, Bulgaria: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-2710.

Berez-Kroeker, Andrea L, Lauren Gawne, Susan Smythe Kung, Barbara F Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. 'Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field'. *De Gruyter*, Linguistics, 56 (1): 18. https://doi.org/10.1515/ling-2017-0032.

Bowern, Claire. 2008. *Linguistic Fieldwork - A Practical Guide*. New York: Palgrave MacMillan.

Haig, Geoffrey & Stefan Schnell. 2016. The discourse basis of ergativity revisited. *Language* 92(3). 591–618. doi:10.1353/lan.2016.0049.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne & Nora C. England. 1992. Endangered languages. *Language* 68(1). 1–42. https://doi.org/10.1353/lan.1992.0052.

Holton, Gary & Mandana Seyfeddinipur. 2018. Reflections on funding to support documentary linguistics. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on Language Documentation 20 Years After Himmelmann 1998* (Language Documentation & Conservation Special Publication 15), 100–109. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24812 (17 December, 2019).

Meakins, Felicity, Jennifer Green & Myfany Turpin. 2018. *Understanding Linguistic Fieldwork*. London, New York: Routledge.

Mettouchi, Amina, Zygmunt Frajzyngier & Christian Chanard (eds.). 2017. *Corpus-based cross-linguistic studies on Predication (CorTypo)*. http://cortypo.huma-num.fr/Publication (3 November, 2018).

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*, 67–85. Berlin: Mouton de Gruyter.

Norcliffe, Elisabeth, Alice C. Harris & T. Florian Jaeger. 2015. Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition and Neuroscience* 30(9). 1009–1032. https://doi.org/10.1080/23273798.2015.1080373.

Stave, Matthew, Ludger Paschen, François Pellegrino & Frank Seifart. 2020. Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. To appear in *Linguistic Vanguard*.

Thieberger, Nicholas & Andrea L. Berez. 2011. Linguistic Data Management. In Nicholas Thieberger (ed.), *The Oxford Handbook of Linguistic Fieldwork*, 90–118. Oxford, New York: Oxford University Press.

Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. Australian Journal of Linguistics 36(1). 1–21. https://doi.org/10.1080/07268602.2016.1109428.

Vasile, Aurelia, Séverine Guillaume, Mourad Aouini & Alexis Michaud. 2020. *Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger*. I2D - Information, données & documents. https://halshs.archives-ouvertes.fr/halshs-02870206.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1). 1–9. https://doi.org/10.1038/sdata.2016.18.